DOCUMENT RESUME

ED 068 115 LI 003 910

AUTHOR	Donaldson, Theodore
TITLE	An Information System for Educational Management, Vol 3: Data Requirements for Evaluation: A Review of Educational Research.
INSTITUTION	Rand Corp., Santa Monica, Calif.
SPONS AGENCY	Los Angeles Unified School District, Calif.
REPORT NO	R-932-LACS
PUB DATE	Dec 71
NOTE	73p.; (106 References)
AVAILABLE FROM	The Rand Corp., 1700 Main St., Santa Monica, Calif. 90406 ($3.00)
EDRS PRICE	MF-$0.65 HC Not Available from EDRS.
DESCRIPTORS	*Data; Decision Making; Education; *Educational Accountability; *Educational Administration; Educational Programs; *Educational Research; Evaluation; Evaluation Criteria; Information Needs; *Information Systems; Management Information Systems
IDENTIFIERS	LAUSD; *Los Angeles Unified School District

ABSTRACT

The determination of what data are available for evaluation and accountability, and what further data are needed are the aims of the educational research survey reported. The results are presented in four chapters dealing respectively with Measures of Educational Outcome, Teacher Effects, Instructional Effects, and Student Characteristics. Inasmuch as schools and innovated education programs are being evaluated by standardized achievement tests, a crucial and immediate need exists to improve test design, concept, scoring, and administration. When student achievement is applied as a criterion of teacher effectiveness, some teacher classroom behavior appears to be consistently associated with better student achievement. Studies of instructional methods used in classroom and curriculum design have produced few consistent and positive results. Findings indicate that, to be effective, educators must develop methods tailored for individual ability. The research surveyed here carries implications for evaluation and accountability, and for designing data systems to support them. (Other reports in this series are: LI003908 and 003909, and LI003911 and 003912). (Author/NH)

VOLUME 3

R-932/LACS
DECEMBER 1971

AN INFORMATION SYSTEM FOR EDUCATIONAL MANAGEMENT: Data Requirements for Evaluation; A Review of Educational Research

Theodore Donaldson

Prepared for the Los Angeles Unified School District

ED 068115

U.S. DEPARTMENT OF HEALTH
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

LI 003 910

Rand
SANTA MONICA, CA. 90406

This report was prepared for the Los Angeles Unified School District under Contract 2410. Views or conclusions contained in this study should not be interpreted as representing the official opinion of LAUSD.

Copyright © 1971
Published by The Rand Corporation

3

VOLUME III
DECEMBER 1971
R-932/LACS

Data Requirements for Evaluation; A Review of Educational Research

Theodore Donaldson

Prepared for the Los Angeles Unified School District

Rand
SANTA MONICA, CA. 90406

PREFACE

In response to community, legislative, and student pressures, school administrators have recently begun to examine the potential of modern management tools and practices. This search for techniques that might function effectively in an educational context led to the adaptation of such methods as program budgeting and accountability. Another tool frequently chosen for educational assistance is the modern management information system, a (usually) computer-based aid to planning and decisionmaking.

In late 1970, the Los Angeles Unified School District (LAUSD) contracted with The Rand Corporation to design such an information system in support of educational management. The system is specifically intended to aid planning and decisionmaking (through implementation of accountability and program budgeting) in schools partially supported by Title I provisions of the Elementary and Secondary Education Act.

This Report is the third in a series describing the proposed information system. In addition to Title I funding, the present Report was supported in part by the President's Commission on School Finance; much of the material also appears in Rand's report to the Commission. The Report reviews and discusses the literature concerning student evaluation, providing direction for eventual information system growth. It also discusses the implications of research findings for accountability. Because the review is lengthy and involved, some readers may wish to skip the full descriptions and, instead, read the summary and the implications of the findings given at the end of each section.

The series also includes:

- J. A. Farquhar and B. W. Boehm, *An Information System for Educational Management, Vol. I: Design Considerations,* R-930-LACS. Defines near-term information system requirements, design guidelines, major design constraints, and information needs of educational decisionmakers.
- M. L. Rapp, *An Information System for Educational Management, Vol. 2: Data Requirements for Accountability,* R-931-LACS. Defines the future shape of the accountability system and feasible long-term trends and requirements in the areas of research and evaluation.

- J. A. Farquhar, D. H. Stewart, J. Lombaerde, *An Information System for Educational Management, Vol. IV: Functional Design*, R-933-LACS. A functional description of the proposed information system, specifying input and output data, file formats, and necessary processing.
- J. A. Farquhar, I. M. Iwashita, S. H. Landa, *An Information System for Educational Management, Vol. V: A Design for Implementation*, R-934-LACS. Describes and discusses alternative hardware, software, and support configurations that might provide the desired services, and the costs and benefits of each.
- L. A. Dougharty and S. A. Haggart, *An Information System for Educational Management, Vol. VI: An In-Service Training Program*, R-935-LACS. Describes the education and training requirements for educational administrators charged with effective use of program budgeting, accountability, and the design information system.

SUMMARY

This Report surveys educational research to determine what data are available for evaluation and accountability, and what further data are needed. The results are presented in four chapters dealing respectively with Measures of Educational Outcome, Teacher Effects, Instructional Effects, and Student Characteristics.

Although standardized achievement tests are widely used to measure educational outcomes for individual students, the tests are generally inadequate because, at best, they quantify only limited aspects of cognitive performance, and higher cognitive abilities and achievements go untested. Inasmuch as schools and innovated education programs are being evaluated in terms of such achievement tests, a crucial and immediate need exists to improve test design, concept, scoring, and administration. Factors in noncognitive achievement are often discussed, and, although important, attempts to measure them are relatively rare.

Until recently, research on teacher effectiveness has not used student achievement as a criterion. When this criterion is applied, some teacher classroom behavior appears to be consistently associated with better student achievement. Studies over time, and/or subject areas, however, indicate little stability in teacher effectiveness. This might be caused by the transitory nature of teacher skills and other characteristics, or by uncontrolled differences in comparative groups and the effect of student-teacher interaction. Some evidence indicates that teacher expectations influence student achievement and behavior; for example, teachers are differentially effective with students, and perhaps classroom performance could be improved by assigning students to teachers on the basis of their ability to work together.

Studies of instructional methods used in classroom and curriculum design have produced few consistent and positive results. Although television learning, teaching machines, and programmed instruction can reach more students, no evidence to date has proven their general superiority over conventional methods. Psychological studies of factors affecting instruction generally use very different learning tasks from those required in the classroom so that the results have limited value for measuring the effectiveness of instructional methods. A promising research area is transfer of learning, especially as it relates to organization of instructional material, which apparently can be structured in a hierarchy, although the rationale and basis

for structuring are not clear. More research is needed on almost every factor being studied.

Educational research has generally failed to consider the student's unique characteristics. Even more, it has typically ignored the interaction of individual characteristics with instructional methods, teacher characteristics, and type of learning task. Review of psychological research and a few education studies shows the importance of individual characteristics. Although no conclusive evidence has been generated for an interaction between any special abilities and education, general intelligence is substantially correlated with the ability to learn, especially for abstract and complex material. In addition, evidence strongly suggests an interaction between intelligence and educational treatment (instruction, task, etc.). Findings indicate that to be effective, educators must develop methods tailored for individual ability. Previous attempts to do this via ability-grouping may have failed because the programs did not fit the unique abilities of each group.

Less definitive research findings indicate that many other factors differentiate students and their response to specific education programs. For example, creativity is not highly dependent on intelligence, nor does high intelligence guarantee creativity; but it appears that the creative person requires a different educational approach than the less creative person. More generally, there are individual differences in understanding concepts which carry important implications for instructional design.

Many personality variables (need for achievement, autonomy, and anxiety, among others) appear to influence school achievement and to interact with educational factors. Because such noncognitive factors apparently affect school achievement, there is a growing interest in the effect that preschool years have on educational outcomes. A number of experimental studies and preschool education programs indicate that major determinants of achievement are formulated during these years.

The research surveyed here carries implications for evaluation and accountability, and for designing data systems to support them. At present, a rigorous approach to accountability appears unfeasible because too little is known about factors that determine educational outcomes. Moreover, the state of the art of education measurement needs further development before it will be possible to develop a precise accountability system.

CONTENTS

PREFACE iii

SUMMARY v

Section

I. INTRODUCTION 1
- Information Sources 2
- Plan of Presentation 2

II. MEASURING EDUCATIONAL OUTCOMES 4
- Introduction 4
- Teacher Grades 5
- Essay Examinations 5
- Noncognitive Achievement 5
- Standardized Achievement Tests 6
  - Derivation of Normative Scores 8
  - Cultural Bias 9
  - Educational Objectives and Test Content 10
  - Test Validity 10
  - Criterion Referenced Tests 11
  - General Intelligence Tests 12
  - Some Statistical Problems in Using Achievement Tests 13
- Summary 14
- Implications 14

III. TEACHER EFFECTS 16
- Introduction 16
- Teacher Characteristics and Student Achievement 17
- Teacher Skills and Effectiveness 18
- Teacher Expectations 20

Student-Teacher Interactions 21
Summary 22
Implications 22

IV. INSTRUCTIONAL EFFECTS 23
Introduction 23
Classroom Instruction 23
Curriculum and Classroom Instruction 24
Television and Programmed Instruction 28
Experimental Work in Instruction 30
Transfer of Learning 31
Reinforcement and Feedback 33
Attention Factors in Learning 33
Retention of Learned Material 34
Summary 35
Implications 35

V. STUDENT CHARACTERISTICS 37
Introduction 37
Abilities and General Intelligence 38
Student Characteristics and Programmed Instruction 40
Student Characteristics and Meaningfulness 41
Concept Attainment 42
Personality Differences 43
Early Development and Learning 46
Summary 47
Implications 48

VI. CONCLUSIONS AND DISCUSSION 50

APPENDIX: INTERACTION EFFECTS 55

BIBLIOGRAPHY 59

I. INTRODUCTION

This Report summarizes education research to determine what type of data a future information system user will need for evaluation and accountability.[1] Future requirements are derived from two sources: current research results, which indicate important data not presently in the system; and future research required to resolve existing issues, along with the data necessary for this research. Both objectives require an examination and review of education research to understand how learning takes place and to determine what factors affect educational outcomes. Research in many fields is relevant to understanding the education process. Studies of the teachers' characteristics (skills, behaviors, personality, for instance) are obviously relevant as are studies of teaching methodology. Basic psychological studies of learning are relevant, but results are often not directly applicable to the classroom. Perhaps most important in the long run are psychological studies of learning in instruction, individual differences, child development, and personality, because these studies are beginning to define student characteristics and instructional practices that are important in determining individual educational outcomes.

The educational research surveyed here covers classroom studies as well as relevant psychological (laboratory) studies. Classroom studies have not generally produced highly definitive results, whereas laboratory studies have produced many significant and consistent results, but their relevance for classroom learning is often unclear. In this Report, distinctions between laboratory and classroom studies are based on objectives, learning tasks, and the kind of outcome measures employed.

Classroom studies have the objective of understanding meaningful classroom learning, and involve some measure of educational outcome (e.g., achievement tests, grades, and teacher or supervisor ratings). Laboratory studies usually have more theoretical objectives (e.g., advancing knowledge about psychological phenomena, developing and testing theory, or investigating empirical relationships between psychological variables), and outcome measures are varied and difficult to summarize. Laboratory measures are generally based on the learning or retention of well defined and highly specific responses, and in addition the experimentalist is not primarily concerned with the amount learned, but with how learning takes place

[1] Volume II of this series describes how currently available data are used for accountability.

and what factors affect learning or retention. For example, an experimenter might present both auditory and visual stimuli in pairs to children to see which sense affects learning and retention most. Each child receives the signals until he can recall without error the second stimulus in each pair upon presentation of only the first. Learning would be measured by the number of presentations the child requires to learn the list of stimulus pairs without error. This same test might be applied across age groups to determine whether or not differences exist as a function of age.

INFORMATION SOURCES

Thousands of relevant education studies are published each year. To review them all would have been impossible within the limitations of the contracted effort, and would undoubtedly have resulted in an unreadable report. Fortunately, review articles in the areas of concern here are published yearly and, in addition, books frequently review broad areas of research. Some reviews simply summarize many studies. Consequently, to determine the merits of a particular study, the researcher must go to the original source. Other reviews critique as well as summarize, relating studies to each other and to basic issues in methodology and education. These reviews are easier to read and to comprehend, although the researcher may be swayed by the reviewer's particular orientation. To give the reader of this Report a comprehensive view of the vast body of research, critical review articles have been utilized. In many instances, the original studies were read to check on the reviewer's summary and conclusion, but in general, original references are not cited. In many cases, the same study is critiqued in more than one review. This is especially true for the more important and meaningful studies, and helps in assessing whether or not a review is biased. The quality of various reviews is indicated throughout this Report. In general, a review was rated high if it summarized across studies, and contained an evaluation that did not contradict the author's own appraisal of critical studies. A study was judged good if it was methodologically sound and meaningful in terms of some underlying issue or problem.

PLAN OF PRESENTATION

The Report is organized as follows. Section II discusses measures of educational outcome, with emphasis on the validity and use of standardized tests. This discussion is important because evaluation depends entirely on adequate measures of educational outcome. Section III deals with the general results of research on teacher characteristics, including research that relates teachers' skills, behavior, attitudes, and personality to some measure of student achievement. Section IV presents the results of research on instructional method; some has been conducted in the classroom, but the majority has been done in the psychological laboratory. This is espe-

cially true with respect to studies that report positive results; most classroom studies are at best inconclusive. Section V summarizes the results of research concerned with the interaction between student and education characteristics, revealing the importance of individual attributes and how they can influence achievement. Section V suggests that students respond differently to education factors (teachers and instructional method) depending on their own characteristics. In the author's opinion, these interactions are extremely important. Finally in Sec. VI the findings are discussed in terms of their implications for accountability. The appendix elaborates on interactive effects.

II. MEASURING EDUCATIONAL OUTCOMES

INTRODUCTION

An education system has many functions and outputs. Some relate directly to the student, others hardly involve him at all. For example, the school system must interact with the community and must provide a number of outcomes to please the community. In doing so, the school may sometimes act in ways that seem to operate against desired outcomes for the student. The school also has a political role and must provide outcomes that allow it to compete within a political system for power, money, and position. Whatever importance one assigns to political and social functions, it seems apparent that they are secondary, and perhaps even irrelevant to the school's presumptive primary objective, which is to educate students. This Report uses student learning as the outcome by which to assess a school; no attempt is made to address political and social objectives.

What exactly does student learning mean? The easiest and perhaps the first definition that comes to mind is to interpret learning as the acquisition of knowledge and cognitive skills. In practice, unfortunately, this has mainly been reduced to measuring and testing retention of specific subject matter, and higher cognitive processes (abstract reasoning, problem solving, and creative thinking, among others) are seldom measured (Klein, 1971). Along with the general failure to measure cognitive achievement properly, there is an almost total failure to evaluate and identify "noncognitive achievement"[2] (Bloom, Hastings, and Madaus, 1971, Chap. 10).

Thus, of the many and diverse kinds of student learning, almost all evaluative research is based on a narrow range of cognitive skills as measured by standardized tests. By and large researchers have not dealt with broad measures of student learning nor with the important problem of individual priorities of educational outcomes; however, many of these same researchers who have been unable to resolve this problem analytically frequently discuss the importance of priorities and

[2] This expression is used because it is becoming vogue in education literature, although "achievement" is not the best term to use in this regard. It would be more accurate to talk about noncognitive growth, but debate over terms seems relatively unproductive as long as it is generally understood what the term noncognitive achievement means.

individual differences in priorities. It is becoming increasingly clear that different educational objectives and values exist as well as individual differences in type and level of ability. Thus from the start we must realize that research based on limited measures, and accounting for relatively few objectives, cannot lead to conclusive generalizations about educational outcomes.

Because research is based almost entirely on standardized tests, most of this chapter is directed at problems associated with them. Before discussing standardized tests, some brief comments are made about teacher grades, essay examinations, and noncognitive achievement.

TEACHER GRADES

Teacher grades of student performance are extremely unreliable; they neither correlate with standardized test scores, nor do teachers correlate with each other in grades assigned to the same student (Cronbach, 1971). Teacher grades are greatly influenced by student characteristics such as docility and social class, and criteria vary from teacher to teacher. Grades are further influenced by school policy factors such as "grading on the curve," or community pressure from parents who do not like to see their children fail. The technical problems associated with grades as a subjective rating system are complex but they need not be discussed here. Grades have played almost no part in the research on evaluation of educational outcomes.

ESSAY EXAMINATIONS

Essay examinations are widely used, sometimes because objective tests cannot be designed to measure some criteria of learning. Although such examinations are widely used, and despite their advantage in measuring broad kinds of cognitive ability, they are generally unreliable. Answers to essay questions vary in several dimensions: vocabulary, style, thought, originality, and neatness, for example. Thus a single score is a complex weighted sum of the scores on each dimension. Moreover, since subscores are rarely worked out by the grader, the relative weights vary between graders, for the same grader over time, and depending on the individual situation. In reviewing the research on essay examinations, Coffman (1971) points out that much work is still needed in developing rules for writing and scoring essay questions. None of the research reviewed in this Report uses essay scores as a measure of educational outcomes.

NONCOGNITIVE ACHIEVEMENT

Noncognitive factors include motivation, attitudes, learning styles, social skills, self-awareness, and even such vague but important concepts as happiness and qual-

ity of life. These factors engender two different viewpoints. One view contends that noncognitive factors are important because they are believed to be the major determinant of cognitive achievement; evidence presented later in this Report implies that they undoubtedly are. The other view holds that growth in noncognitive rather than cognitive factors is the more relevant goal of education. These views are certainly not mutually exclusive, and most educators agree that noncognitive factors are important for both reasons. In fact, the distinction between cognitive and noncognitive achievement is rather artificial: attitudes and motivation have strong intrinsic cognitive components, and cognitive skills and abilities have strong intrinsic noncognitive components.

Education in general and compensatory programs in particular are concerned with improving student motivation, attitudes, and general affective (noncognitive) behavior. Generalization of cognitive ability results not only from the transfer of specific skills, but also from such noncognitive factors as the establishment of learning styles, learning sets, motivation for learning, and attitudes about learning. Noncognitive factors undoubtedly outweigh the importance of specific cognitive skills for future learning, although acquiring cognitive skills may itself considerably affect noncognitive factors such as motivation, self-awareness, and the like. In their book on evaluation of learning, Bloom, Hastings, and Madaus (1971) devote an entire chapter to measuring affective behavior, and include affective goals in stated educational objectives. Recent research literature, especially that related to compensatory and preschool education, repeatedly comments on the importance of noncognitive factors in determining cognitive achievement and the necessity of identifying, measuring, and shaping these factors at an early age (e.g., Denenberg, 1970).

Noncognitive factors have even greater significance because of recent evidence showing the low correlation between cognitive achievement (measured by grades and standardized tests) and later life success. Cohen (1970), Gintis (1971), and Holtzman (1971) cite evidence indicating that achievement in terms of job, social class, and general life expectations is apparently only incidentally related to school achievement: it is true that a high correlation exists between *amount* of education and amount of income, but preliminary evidence indicates that the relationship is not usually a causal one. Moreover, Gintis promotes the thesis that noncognitive factors strongly influence worker earnings and productivity. He reviews evidence supporting this thesis, and moreover shows that important dimensions of noncognitive achievement are not promoted or rewarded in most conventional schools. Schools need to include noncognitive factors in their educational objectives, and better methods for evaluating such factors need to be developed.

STANDARDIZED ACHIEVEMENT TESTS

Although the following discusses only the more limited aspects of student achievement, many problems need to be recognized explicitly. With increasing interest in accountability, student achievement is being measured more and more by

standardized tests,[3] with test scores based on national norms. While this practice allows a school to assess itself relative to other schools, the tests introduce a number of liabilities and hazards. Foremost among these is the danger of suppressing desirable outcomes, which are not measured by standardized tests (for example, abstract reasoning, and creativity).

Further, while it certainly is necessary and important for children to acquire basic reading and math skills, focusing on teaching these skills may be less important than is often believed. Generally, achievement in basic math and reading skills as measured by standardized tests is assumed correlated with, and perhaps responsible for, achievement in other subject matters and cognitive areas. However, the generalization[4] of improvement in basic reading and math skills through special programs has not been demonstrated; although in view of the rather temporary nature of many of the gains obtained in these programs, the lack of generalization is not surprising. Undoubtedly, these skills do generalize under some conditions, but the conditions are not known; this is discussed in Sec. IV.

In addition to the general difficulties discussed above, there are a number of specific problems associated with standardized tests. The UCLA Center for the Study of Evaluation reviewed over 1500 standardized tests used in elementary schools (Hoepfner, 1970). Results indicate the tests by and large are unsatisfactory. Klein (1971) has written a strong critique of standardized tests and their misuse and summarized as follows:

> So far, the discussion has painted a pretty bleak picture regarding the utility of standardized tests for accountability. The major problems involve questionable test validity, poor overlap between program and test objectives, inappropriate test instructions and directions, and confusing test designs and formats. In short, a void exists between the demands of accountability and the present stock of standardized instruments. Further, this void will probably only widen as the pressure for accountability increases unless we start improving the methods of test construction and use.

Although Klein's comments are applied to accountability, they are also true for evaluation in general. The first step in accountability is evaluation, and in this respect, achievement tests are generally weak. As Anastasi (1967), among others, has pointed out, improvements are needed more in the interpretation of scores and orientation of users than in the actual construction of test instruments. A number of the technical problems in using these tests are discussed.

[3] The most widely used standardized tests measure achievement in subject matter areas, although there are also many tests for math and reading readiness, concept attainment, psycholinguistic performance, and other general and specific ability tests. In the elementary grades, the most widely used tests are those for math and reading ability, and the current programs of performance contracting and accountability have focused almost entirely on measuring these skills.

[4] Generalization is the spreading of acquired skills to areas in which the student has had no specific practice. For example, generalization (or transfer) occurs when an improvement in basic reading skills leads to (1) an improvement in concurrent school achievement, such as proficiency in social studies or science; and/or (2) an improvement in future school achievement, including reading. In most cases, it is assumed that no special practice takes place other than in reading skills.

Derivation of Normative Scores

To understand many of the problems associated with standardized tests, it is necessary to understand how normative scores are derived. Assuming test items actually measure the amount of learning that has taken place in a course of instruction, normative scores are necessary to determine what a "raw" test score[5] means (cumulated over all items). For example, how much "better" is a raw score of 70 than one of 60 (i.e., how much more about the course does the student know)? How high a score should be "expected"? These questions and others are answered by deriving a normative score from actual test scores.

Essentially, the normative score indicates a student's position in a distribution of scores. To determine the reference distribution, a sample from a specified population is selected and given the test (e.g., 4th grade children in California). A given individual's raw score can then be represented as higher than X percent of the sample scores, or we would say his score is at the Xth percentile. Since the sample distribution is "close" to the population distribution, the percentile score represents the student's position in relation to the general reference population. Percentile scores can be transformed into grade equivalent or other types of normative scores.

Although grade and age[6] equivalent scores are widely used, they have been severely criticized (Cronbach, 1970; Angoff, 1971). Angoff presents a penetrating analysis of the technical and practical problems associated with the use of grade and age equivalent scores. Equivalent scores are obtained by administering a test to samples of children over the range of desired grades (or age). The average for a grade (or 50th percentile score) determines the grade level score. A regression line is then plotted between the mean score obtained by each grade across all grades. This regression is used to determine a child's grade equivalent score by simply noting where his score falls on the regression line. Had the regression of grade on score (rather than score on grade) been used, a different regression line would have resulted, and scores would have different grade equivalents.

This basic ambiguity is further compounded because the equivalent score interpretation depends upon the variation of scores about the mean for each grade in the original sample (i.e., the variation about the regression line). A child who is two grades advanced on a test of high reliability (low variability about the regression line) is also high in his percentile rank (say 95). But with a test of low reliability (high variability), the same two-year advanced status is associated with a much smaller percentile rank (say 70). Further, a 6th grader with a 9th grade equivalent score does not possess a 9th grader's skills, nor is he psychologically the same. Cronbach (1970, p. 98) comments on equivalent scores:

> In the writer's opinion, grade conversions should never be used in reporting on a pupil or a class, or in research. Standard scores or percentiles of raw

[5] A raw score is a measure of the actual number of correct responses. The score may be a simple frequency count or it may be the sum of test points, with each test item given some arbitrary assignment of possible points.

[6] Age equivalents are most often used with mental abilities tests, and they report a "mental" age score. The score represents age level relative to mean performance on an ability on age regression line.

scores serve better. Age conversions are also likely to be misinterpreted. A 6-year-old with mental age 9 cannot pass the tests a 12-year-old with mental age 9 passes; the two simply passed about the same fraction of the test tasks. On the whole, however, age equivalents cause less trouble than grade equivalents, if only because the former are not used for policy decisions in education.

These comments represent only the highlights of the problems inherent in equivalent scores. For a detailed treatise, the reader is referred to Angoff (1971).

Cultural Bias

An important issue in deriving standardized scores concerns the choice of the normative population. A test with national norms is one that is supposedly based on a sample representing the normative population across the nation. To be accurate, the sample population must be stratified in the same proportions as the overall population, i.e., Negroes and Caucasians, poor and rich, must appear in the sample in the same proportion as they appear in the general population. This means that any nationally normed test primarily reflects the characteristics of white, middle-class America, simply because they represent the greatest proportion of the population.

One form of cultural bias arises when a test is normed on one population and used to test people from another population. The bias can be subtle and may lead to gross misinterpretations of data. For example, a nationally normed test of concept ability might be given to children from a Mexican-American ghetto. If the test uses written test items and instructions, the children's scores are affected by their ability to understand language; and if they have language problems, their concept ability scores will be poor. Their "true" concept ability remains untested. Attempts to develop tests that are free from language ability have not been very successful; even "nonverbal" tests are frequently found to correlate with language ability.

A more subtle influence of the normative population occurs through the operation of its values. Due to the way standardized tests are constructed, they necessarily reflect what the normative population feels is important. Without great exaggeration, one may state that these tests indicate how well students have achieved white, middle-class goals. Later we quote a comment by Jensen illustrating this point in reference to intelligence tests. Holtzman (1971, p. 551) discusses the problem of cultural bias in testing and emergency social issues:

> The emergence of black culture, the Chicano movement, and the stirring of the American Indian as well as other forgotten groups in the wake of desegregation and civil rights legislation have forced white America to re-examine its soul. The results in the field of mental measurement have been a recognition and acceptance of cultural variability, a search for new kinds of cognitive, perceptual, and affective measures by which to gauge mental development, and a renewed determination to contribute significantly to the task of overcoming educational and intellectual deprivation.

In general, tests designed for normative use discriminate against those who are culturally different from the majority.

Educational Objectives and Test Content

The apparent failure of many innovative educational programs is often said to occur because standardized tests used to evaluate the programs do not measure outcome in terms of some or all program objectives (e.g., Cohen, 1970; Klein, 1971; Lennon, 1971). Part of the problem is that objectives are rarely stated with sufficient clarity; but even overlooking this liability, the match between program and test objectives is often poor. In the first place, as Klein (1971) points out, valid tests covering all of the objectives a school might like to attain do not exist.

Second, tests may cover some program objectives, but there is usually poor agreement between the specific objectives and the test content. For example, a test may measure reading ability in terms of, say, eight areas. A specific program might be aimed at only six objectives, with no interest in the other two. Most tests, however, only report a *single* score averaged across all areas, and this score indicates achievement on all eight objectives. So a score would be a combination of how well a student achieved on the six reading program objectives, plus how well he achieved on the other two. This makes it impossible to evaluate the program objectives. Tests are not designed with specific programs in mind, and poor overlap is to be expected between the objectives a test measures and those an education program aspires to. Another complication occurs when the test does not represent test objectives equally. Some of these problems would be clarified if the tests reported separate scores for each area or objective.

Stating and evaluating educational objectives may be one of the most crucial problems in educational research. Undoubtedly much of the evidence for lack of improvement in education derives from evaluation which does not cover all program objectives.

Test Validity

Test validity generally means "does the test measure what it is supposed to measure?" and is formally determined by a number of techniques.[7] One, a complex process called *construct* validity, essentially determines how highly tests supposedly measuring the same thing correlate with each other. Lack of high correlation indicates that one or all of the tests do not validly measure the construct being considered. A second kind of validity is called *predictive*, and in this procedure the test is correlated with an external criterion. For example, a test of reading readiness might be validated by using success in a reading course as a criterion. The assumption is that better readiness leads to better achievement. In practice, both kinds of measures are necessary for test validity. A third type, sometimes referred to as *face*

[7] For a detailed discussion, see Cronbach (1970).

validity simply asks if the items in the test appear to measure what the test is designed to measure. While this latter method lacks the sophistication of the first two, many standardized tests even fail on this measure. Klein (1971) points out several examples in which it is obvious that the test items have little to do with what the test purports to measure. There are, however, many tests that are purposely designed without consideration of face validity, although they are not widely used in education. Finally, a test is said to have *content* validity if some authority asserts that the test measures something that it purports to measure. Much of the foregoing discussion on the relationship of objectives to test content relates to content validity. The four measures of validity are all methods for determining the same thing, and generally several methods are used in determining the authenticity of a given test.

As previously mentioned, tests often do not adequately overlap program objectives, and generally they are not valid even when they do appear to overlap. Bormuth (1970), in a book on the theory and design of test items, criticizes current methods of test construction on the grounds that the item generation techniques lead to tests of low validity. An item represents the test writer's response to instructional material, and the student's score is thus a function of the test writer and has no known relationship to instructional content. Bormuth goes on to develop a method for deriving test items based on a linguistic analysis of the instructional content and the instruction objectives.

Criterion Referenced Tests

Standardized (normative) tests are sometimes criticized because their scores do not indicate the specific skills a student masters. They only place him relative to other students, and not relative to instructional content. For example, two students scoring at the fiftieth percentile on a reading test could have answered different questions correctly and have acquired different reading skills. This is true even if the test gives percentile scores for a number of subskills; they are still normative scores. This problem is being attacked through the design of so-called *criterion referenced tests* (Cronbach, 1970; Glaser and Nitko, 1971). Each item on a criterion referenced test is designed to measure or indicate the accomplishment of a particular skill. The number of items passed is not the important factor, but rather which items are passed. The student is not allowed to proceed to advanced instruction until acquiring prerequisite knowledge.

A key feature of criterion referenced tests is their relationship to the specific goals and subject matter of a course. Test items are designed to indicate success on the learning tasks necessary to cover the subject matter and to meet the course objectives. This requires a detailed task analysis of course material. Few procedures for the task analysis have been developed, although Gagné's work on hierarchical organization shows promise. Section IV discusses research on the organization of instructional material, and there we point out that skills and knowledge required for a course can be arranged in a hierarchy, such that success at a higher level depends upon the acquisition of skills at a lower level.

The distinction between normative and criterion referenced tests is made primarily on the basis of the purpose for which the test was constructed and how information obtained from it is used. The purpose of a criterion referenced test is to indicate a student's status on a set of specific tasks necessary for completing a course of instruction. The test information not only assesses his accomplishments but is also used to determine what tasks the student is ready to undertake. Norm referenced tests indicate a student's relative position in a population, and the information from these tests is used to evaluate achievement relative to other students. This summary score can be used for comparing students, or groups of students, in terms of overall achievement. The use of criterion referenced tests for this purpose is not clear since such tests indicate which instructional tasks the student has accomplished; essentially, he passes or he does not for each task. The *number* of tasks he "passes" cannot be meaningfully added for a total test score. Criterion referenced tests serve diagnostic functions in evaluation, which aims at special information for student remediation or course improvement.

Much work remains to be done in developing criterion referenced tests, but they appear to have great promise. Their greatest potential and value are that these tests focus on instructional content, yield information for remediation, and allow for individual differences in performance.

General Intelligence Tests

General intelligence tests are standardized achievement tests. They have been developed over a longer period than most standardized achievement tests, and more research has been directed toward their improvement; they are more valid when properly used; they usually report subscores on various test objectives; and directions for administration are generally better. Sometimes changes in IQ scores are used to measure student achievement, and many attempts have been made to improve IQ scores through compensatory school and preschool programs. Failure to find consistent evidence that IQ can be modified (e.g., Butler, 1970) led Kohlberg (1968), among others, to argue that IQ is not a good measure of the efficacy of these programs. For years, psychologists have stated that many IQ tests basically measure achievement. They measure what the person has learned, not primarily his capacity for learning. The scores reflect environmental influences and past learning as well as innate ability. The belief that IQ can be affected by environment has been confirmed many times in studies of identical twins, but many factors contribute to this effect other than those present in the school environment (Vandenberg, 1966). On the other hand, Jensen (1969) reports evidence that IQ is largely determined by genetics, and can only be modified by environment in a relatively small degree.

The various uses of IQ tests in recent education programs has caused a reemergence of debate and inquiry into the validity and meaning of general intelligence test scores. The crucial factor in determining the appropriateness of their use (or any achievement test) depends on the goals and objectives the test is being used to evaluate. This is never an easy task, and is made even more difficult by the

interaction of social values and subtle and nonverbalized goals that exert profound influence on test content, scores, and interpretation. This has been well stated by Jensen (1970):

> ... It should not be forgotten that intelligence tests as we know them evolved in close conjunction with the educational curricula and instructional methods for Europe and North America. Schooling was not simply invented in a single stroke. It has a long evolutionary history and still heavily bears the imprint of its origins in predominantly aristocratic and upper-class European society. Not only did the content of education help to shape this society, but, even more, the nature of the society shaped the content of education and the methods of instruction for imparting it. If the educational needs and goals of this upper segment of society had been different, and if their modal pattern of abilities—both innate abilities and those acquired in these peculiar environmental circumstances—were different, it seems a safe conjecture that the evaluation of educational content and practices and consequently the character of public education in modern times would be quite different from what it is. And our intelligence tests—assuming we have them under these different conditions—would most likely also have taken on a different character.

Some Statistical Problems in Using Achievement Tests

Inadequacies in the use of achievement test scores in education evaluation are partly attributable to the frequent use of faulty statistical analyses. By far the majority of studies on compensatory programs report data on achievement gain[8] over some period, and performance contract agreements are almost exclusively written in terms of achievement gain scores. Gain scores are extremely biased estimates of true gain (for example, see Harris, 1963). An article by Cronbach and Furby (1970) offers some refinements on techniques for estimating true score; however, the important message is that the authors see no advantage to using gain scores in the first place. Status scores (scores at any point in time) contain all the information given in change scores, at least for the situations in which change scores have traditionally been used. For example, if it is necessary to evaluate the improvement produced by an innovative program, this is best accomplished using a control group. In both treatment and control groups, only the final status or achievement score need be used. Pre-test scores can be involved in the statistical analyses, but not in computing gains. The groups are not compared with respect to each other. In many instances, it is unnecessary to actually use an experimental control group; instead it is possible to use the past history of the system as a benchmark.

[8] A student's test performance is determined by many factors other than his "true" knowledge or ability. Because these other factors vary over time, a person's test score will also vary, so that any given test score is an estimate of the true state of his knowledge or ability. The achieved test score may be a percentile, and age equivalent, or a simple sum of correct items. A gain score is obtained by subtracting the scores obtained by the student on two occasions.

While problems of statistical sophistication and reliability are important, it would seem that the crucial problems in achievement evaluation are not primarily statistical. We agree with Klein (1971) and others that there needs to be a rather complete overhauling of testing procedures and interpetation. The shortcomings of standardized tests must be accounted for in evaluating education. Efforts to eliminate these inadequacies for future evaluation work will require substantial research.

SUMMARY

Using standardized tests to evaluate student achievement has become a major enterprise in the schools; but in spite of the wide use and reliance on these tests, they are generally inadequate. This is alarming in light of the growing activity in evaluation of education outcome based on standardized test scores. At best, generally used tests measure only limited aspects of cognitive performance, while higher cognitive abilities and achievements go untested. Noncognitive achievement is sometimes talked about, but the evaluation of these factors is still in a very crude state. Inasmuch as schools and innovative education programs are being evaluated in terms of such limitations, there is a need for immediate improvements in test design, concept, scoring, interpretation, and administration.

IMPLICATIONS

Currently, the fundamental measure of educational outcomes for accountability is student achievement as measured by standardized achievement tests. Although these scores appear more reliable than teacher grades, they are not good indicators of student achievement even in a limited sense. For accountability to work, these tests must be improved; the scope of achievements they measure must especially be extended. In the meantime, a great deal of caution is necessary in interpreting results based on these data.

This raises the question of what information can be used to assess educational outcomes. Objectives of education are broad, therefore it is necessary to obtain data on student performance over a broad and relevant (to objectives) range of school performance. As an experimental program based on a relatively small sample, the technique suggested by Donaldson (1971) for subjectively scaling student performance in school work holds promise. This method allows the user to evaluate student performance, and these data could serve as a criterion for further evaluation of standardized tests. Analysis of the patterns of discrepancies with the standardized scores would make possible a diagnosis of the limitations and inadequacies of both procedures.

In view of the fallibility of achievement test scores and teacher grades, data

systems for accountability cannot become dependent on such information, and the system must be able and willing to incorporate new data sources. There is need for much research on adequate data for accountability. In the meantime, selection of tests and interpretation of test data should be pursued cautiously, and if possible, experts (such as the Center for Study of Evaluation at UCLA) should be consulted.

III. TEACHER EFFECTS

INTRODUCTION

Studies of teacher characteristics have abounded since the 1930s, and now number in the thousands. Despite this, little is known about what constitutes desirable teacher characteristics, and especially what their influence is on student performance. With the exception of a few recent studies, the use of student achievement as a criterion to evaluate teacher performance has rarely occurred, and therein lies a great weakness. Attempts to use criteria such as supervisor or fellow teacher ratings are not successful in that the ratings do not correlate with student achievement (Harris, 1969). This could mean either that the ratings are based on indicators of success other than achievement, or that supervisors and teachers do not have a good idea of what constitutes superior teaching. Of course, this may also be another example of the influence of a student-teacher interaction, so that overall effects are difficult to isolate.

Past research has focused almost entirely on measuring various teacher attitudes and personality traits, with some attempts to relate these to supervisors' estimates of classroom success. More often, the studies simply intercorrelate various tests of teacher attitudes, interests, intelligence, and so forth. In the end, the studies are either contradictory or have little practical value, and often have both problems. To quote Getzel and Jackson (1963, p. 574):

> For example, it is said after the usual inventory tabulation that good teachers are friendly, cheerful, sympathetic, and morally virtuous rather than cruel, depressed, unsympathetic, and morally depraved. But when this has been said, not very much that is especially useful has been revealed. For what conceivable human interaction—and teaching implies first and foremost a human interaction—is not the better if the people involved are friendly, cheerful, sympathetic, and virtuous rather than the opposite?

In any event, there is reason for skepticism concerning the payoff in studies of teacher attitudes and personality characteristics. Variables related to attitude and personality are difficult to define and more difficult to measure, especially in what

is essentially a normal (healthy) population. Further, it seems reasonable to assume that teacher classroom behavior and technique are more important than attitude or personality. Of course, dimensions of attitude and personality are reflected in the teacher's classroom behavior (Turner and Denny, 1969), and particularly in the degree to which the behavior can or cannot be modified through training. Whatever the influence of personality and attitude factors, however, it is the teacher's classroom behavior that the student responds to, and it is necessary to understand how this behavior is related to student achievement.

TEACHER CHARACTERISTICS AND STUDENT ACHIEVEMENT

In exception to the bulk of research on teacher characteristics, there are a few recent studies, ten of them experimental and fifty correlational, that relate teacher classroom behavior to student achievement. These are described later in this section. There are two general methods used for studying the effects of teacher behavior on student achievement. The best approach is *experimental,* in which teachers are trained in a specific method and student achievement under this method is contrasted with student achievement under an alternative technique.

Studies of this type must meet all the demands of an experimental approach (e.g., random assignment of students to teachers), plus some special demands arising from the situation. Foremost among these special demands is the requirement for measures of classroom transactions, since only by observing the teacher is it possible to determine whether the intended method is actually used. In addition, data on classroom transactions are the only source of information on the content (rather than result) of the student-teacher relationship. Many studies are lacking in such measures, and therefore studies of different teaching methods are rendered useless. In an excellent review of research on teaching, Rosenshine and Furst (1971) could find no more than ten studies that use the experimental method adequately and that provide data on classroom transactions.

A more frequently used procedure for relating teacher performance to student achievement is to *correlate* the two as they occur in the normal classroom. That is, no attempt is made to manipulate teaching methods experimentally. Various dimensions of teacher behavior are observed and rated, and the ratings are correlated with some dimension of student achievement. This approach is dangerous in that correlational relationships suggest causative connections. For example, a high correlation between clarity of presentation and student achievemenet does not mean that clarity *causes* high achievement. It is just as likely that both result from some other factor, say teacher verbal ability or general intelligence. Rosenshine and Furst (1971) find approximately fifty studies that use this procedure.

Studies using the experimental and correlational approaches have produced some consistent and significant results. Rosenshine and Furst summarize these and group them according to eleven kinds of behavior that interrelate significantly with achievement scores. The research strongly supports five of these: clarity of teacher

presentation; variability of teacher classroom activities; teacher enthusiasm; degree to which the teacher was task- or achievement-oriented and/or businesslike; and student opportunity to learn criterion material. The other six variables that were less related to student achievement are the following: use of student ideas and/or teacher indirectness; use of criticism; use of structuring comments; use of multiple levels of discourse; probing; and perceived difficulty of the course. In summarizing non-significant results the authors note:

> At first glance, the above list of the strongest findings may appear to represent mere educational platitudes. Their value can be appreciated, however, only when they are compared to the behavioral characteristics, equally virtuous and "obvious," which have not shown significant or consistent relationships with achievement *to date*. These variables . . . are listed below, and the method by which they were assessed follows in parenthesis: nonverbal approval (counting), praise (counting), warmth (rating), ratio of all indirect behaviors to all direct teacher behaviors, or the I/D ratio (counting), flexibility (counting), questions or interchanges classified into two types (counting), teacher talk (counting), student talk (counting), student participation (rating), number of teacher-student interactions (counting), student absence, teacher absence, teacher time spent on class participation (rating), teacher experience, and teacher knowledge of subject area.[9]

The authors go on to discuss refinements that are necessary in future correlational studies. Of great importance is the need for more experimentally controlled research, with better measures of classroom transaction and broad indicators of outcome measures of student achievement. Classroom effectiveness studies of teacher and instructional techniques depend on the refinement and increased use of observational data systems. Many articles comment on this need, and there have been a number of attempts to develop and refine observational data systems (i.e., Bloom, et al., 1971; Rosenshine, 1970a; Hanley, 1970). Unfortunately, none have been used widely enough or consistently enough to fully realize their potential.

TEACHER SKILLS AND EFFECTIVENESS

The teacher's classroom skills are obviously an important factor in determining educational outcomes. These skills are rarely determined directly, however, and most investigations simply rely on supervisor's ratings. The only studies found that measured teacher skills directly were by Turner (1968), who investigated differences in teacher skills and characteristics as a function of school district characteristics.

[9] (Rosenshine and Furst, 1971.) Counting refers to the number of times a specified behavior occurred. Rating refers to subjective estimates by a judge (teacher, student, observer) regarding teacher performance with respect to some behavior. The behavior is rated into a number of categories in terms of desirability.

In this and previous studies, he developed instruments for measuring the effectiveness of teacher presentation of cognitive material in the classroom. These instruments measure teacher skills in diagnosing learning difficulties and in organizing or sequencing learning material in reading, arithmetic, and science. The study also includes measures of teacher personal-social factors encompassing warmth-spontaneity, classroom organization, educational viewpoint, emotional stability and involvement. The validity of the various scales were determined by measures of internal consistency, i.e., the degree to which teachers scored consistently on each scale. *Validity was never determined on the basis of a relationship to student achievement.* Study results indicate that teachers differ significantly in these characteristics, and that a relationship exists between attractiveness of school districts and teacher characteristics (which does not seem terribly surprising). Before making much of these results, the teacher characteristics must be related to student performance. It is perhaps important to know that attractive school districts (in terms of location, money and students) obtain teachers who apparently have the more desirable characteristics. The important question is do these characteristics make a difference in student achievement, and further for what kinds of students do they make a difference?

In a later study, Turner and Denny (1969) relate the above-mentioned teacher characteristics to student creativity as measured by a scale Denny and others developed. In summarizing, the authors state (p. 209):

> . . . teacher characteristics are distinctly associated with changes in pupil characteristics, as well as with teachers' behaviors in the classroom, which in turn are associated with changes in pupil characteristics. Specifically, the results reported suggest that teachers characterized as warm and spontaneous and teachers characterized as child-centered tend to obtain the greater positive changes in pupil-creativity. These changes appear to come about through teacher classroom behaviors that involve positive reinforcement of pupil responses, through adaptation of activities to pupils, through attention to individuals, and through variation in activities and materials.

Unfortunately, the authors do not present their procedures or data in sufficient detail to allow evaluation of the study. But if the results can be replicated, the findings and method used are certainly important. For one thing, a measure of student outcome other than cognitive achievement was used, although the results would be stronger if a measure of cognitive achievement had *also* been used.

If teachers vary significantly in their skills and classroom behavior, one would expect differences in teacher effectiveness as indicated by student achievement. Rosenshine (1970b) provides a summary and critical review of nine studies of teacher effectiveness. He describes four studies of long-term effectiveness, three of which measured effectiveness over a school year and used grade school teachers.

All of the studies that investigated long-term effectiveness were based on teaching the same material to different students. The three studies of interest used standardized achievement tests which give a number of subtest scores in various

abilities or achievements (Stanford Reading Test, Metropolitan Achievement Tests, among others). The correlations between the mean of a group of students and teachers were generally around 0.35 or much lower, with one study showing a correlation of about 0.50 for two out of five subtests. The results indicate that teachers are not generally stable in their teaching effectiveness of the same material over ti

The other five studies that Rosenshine reviewed concerned short-term effectiveness, with teaching sessions of thirty minutes or less, in which teachers taught (1) the same topic to different groups of students (three studies), (2) different topics to the same group of students (four studies), or (3) different topics to different groups of students (four studies). In each case the same investigator (Fortune) carried out three of the studies. Students were drawn from Head Start to the twelfth grade. When teachers taught the same topic to different students (case 1), the correlations between student groups and teachers were moderate (0.22 to 0.70; but in cases 2 and 3 the correlations were extremely erratic and few were significant.

Such data raise doubts about the meaningfulness of Turner's findings. Although teachers may vary in skill, their effectiveness may not generalize over time or topics. Studies of both teacher skills and effectiveness are extremely limited, however, and any conclusion must be tentative. In addition, while it is necessary to relate teacher skills and characteristics to student achievement, there are certainly grounds for questioning the adequacy of student achievement measures used in these studies. Teachers may be consistent in their effectiveness on other dimensions of educational outcomes, but apparently there are no studies describing this possibility.

The apparent lack of stability in teacher effectiveness may explain in part why these studies of teacher characteristics have proven so futile—their characteristics have no consistent effect, or the characteristics are unstable. Finally, the low correlations may result from a student-teacher-subject interaction. Teachers are not equally effective with all students and all topics, and correlations vary with topic and specific student characteristics.

TEACHER EXPECTATIONS

Previous research (Rosenthal and Jacobsen, 1968) described the importance of teacher expectations as a determinant of student performance. But this report has been criticized on methodological grounds (Snow, 1969) and few of the effects reported appear to be substantial. Such studies are further criticized because they lack data on causative factors, either in establishing teacher expectations, or in terms of mechanisms by which the teacher communicates these expectancies. Two recent studies (Rist, 1960; and Brophy and Good, 1971) investigate some of the mechanisms involved in the establishment, communication, and effect of teacher expectations.

Rist attempted to uncover factors that establish the teacher's expectations about the student, and the effect that such expectations have on the classroom

behavior of both teacher and student. The study followed a single class of ghetto children through kindergarten, first, and second grades. He indicates that in kindergarten, the teacher's expectations and identification of students as "fast" or "slow" learners are essentially based on social class membership.

Brophy and Good investigated how teachers communicate their expectations to first-grade children. Expectations were determined by the teacher's rating of students, but no information was gained about how expectations are established. Apparently, teachers demand better performance from children whom they rate high in their expectations, and praise them when good performance is elicited. They demand less from those whom they expect less from, and tend to withhold praise when good performance occurs.

A few other studies have attempted to verify the effect of teacher expectation. In general it appears that teachers' expectations probably influence teacher and student behavior and may influence measured student achievement. More research is needed to follow up the interesting hypothesis of the "self-fulfilling prophecy."

STUDENT-TEACHER INTERACTIONS

Thelen (1967) reports on direct evidence of the interaction between students and teachers such that some teachers are better with some students than with others, and outlines a method for using this phenomenon to improve classroom behavior and outcome along a number of dimensions. Essentially the method involves assigning students to teachers on the basis of the kind of student the teacher works with best. The method begins with the teacher identifying students he believes are "getting a lot out of class" versus those "not getting a lot out of class." The teacher does not describe these students in any way, but simply points them out. Teachers do not tend to assign the same students to the two categories, and Thelen notes (p. 189):

> Finally, we found that teachers recognize four kinds of students: good, bad, indifferent, and sick. But the problem is that each teacher places different students in these categories, so that whatever is being judged is certainly *not* primarily some characteristic of the student.

The method then establishes the characteristics of students placed in the two categories. In assigning students to teachers two criteria can be used: (1) teachers are given students they work most effectively with, or (2) students are assigned to teachers they can learn from most effectively. This requires determining the kinds of students that have high achievement (relative to their own performance) with a teacher, and then assigning him students of this type. Thelen's study indicates that the same student-teacher grouping would not necessarily result from applying these two criteria, although there would be considerable overlap. In any case, the students are better off when assigned to teachers by either criterion, rather than by just

being arbitrarily assigned to a teacher.

It follows that not only do some teachers do better with some students, but also that there is no single "best" or "right" way to teach. Future research must account for the different teacher preferences and abilities. It makes little sense to talk about teacher skills without also considering the population of students best suited for those skills. Studies of long-term trends in teacher effectiveness must designate which kinds of students the teacher is effective with as well as how effective he is.

SUMMARY

Research on teacher characteristics has generally been extremely uninformative, largely because until recently there have been few attempts to use student achievement as a criterion. Studies of teacher personality and attitudes have produced little, and in view of test instruments for these factors, their future prospects are also poor. Teacher skills and classroom behaviors are measurable and thus show promise, and there appear to be some teacher classroom behaviors that are consistently associated with better student achievement. Studies of teacher effectiveness over time and/or across subject areas indicate very little stability in teacher effectiveness. This could result from a transitory nature in teacher skills and other characteristics, or it could result from uncontrolled differences in comparative groups of the student-teacher interaction. The teacher's expectation about the student has been found to affect student achievement, although these findings are not as firm as some authors would have us believe. Teachers are differentially effective with students, and it appears possible to improve classroom performance by assigning students to teachers on the basis of their ability to work together.

IMPLICATIONS

Research on teachers seems to imply rather strongly that it is fruitless to collect data on personality and attitude factors. If data are desired, the focus should be on teaching skills, classroom behavior and, most of all, on classroom transactions. A continuous record of transaction data is necessary for research but is too expensive to obtain as part of an operating system. Research results seem to have more implication for teacher training and for developing policies for assigning students to teachers than for data systems. The crucial data in evaluating teacher effectiveness are those on student achievement.

IV. INSTRUCTIONAL EFFECTS

INTRODUCTION

To simplify this brief overview of the complex and varied research on instruction, this section is separated into two main subsections. The first looks at instructional methods that are primarily related to classroom learning; the second reviews the psychological research, mostly in the field of learning, that has direct relevance for the design of instructional techniques. Classroom studies investigate school learning. Psychological studies use learning tasks that are dissimilar from normal classroom material; their basic intent is theoretical rather than applied. Psychological studies may be referred to as laboratory studies, although the laboratory may be a classroom.

The distinction between the two, which is somewhat arbitrary at best, is based on the learning tasks that they use rather than where the study occurs. Studies in both sections are the direct outgrowth of the experimental-learning tradition in psychology, and little reference to individual learner characteristics is present.

CLASSROOM INSTRUCTION

Despite years of research on classroom instruction methods, there is little firm evidence to support any particular practice. While this is partially due to inadequate experimental procedures, it probably reflects even more the difficulty and complexity of the problem. Research is an evolutionary process, and although past studies have not produced a substantial body of findings, they have sharpened issues and improved research techniques. By and large, little effort has been expended on identifying student and teacher factors that affect outcome, and even less effort on designing instruments and methods for measuring these factors. Rosenshine (1970a) points out that very few studies contain data on classroom transactions, without which nothing is known about what actually takes place in the classroom. Studies of curriculum design generally suffer from the same problems.

This section begins by briefly summarizing curriculum and instruction re-

search. Curriculum refers to the instructional material and designs for its use. Instruction refers to the interaction between student and teacher as the materials are used. Some general results found for teaching machines, television and programmed instruction then follow.

Curriculum and Classroom Instruction

An enormous amount has been written about curriculum design and use. Westbury (1970) begins a review with the comment:

> Curriculum evaluation appeared as a topic of a chapter in three of five issues of the 1969 *Review of Educational Research*. The emphasis on this topic is, if nothing else, disconcerting to a reviewer who must plow the same field again; it is also puzzling when compared with the infrequent appearances of evaluations of actual curricula or curricular materials in either the research or the subject journals.

and later (p. 239):

> Evaluations exist in the files and reports of those who developed curricula. Yet, while these evaluations remain in files, the proposals and prescriptions of developers circulate freely, without any readily available critical scrutiny. There is a literature of curriculum evaluation, but it is neither publicly available in journals nor has it grown out of an accessible tradition of formal or informal appraisal of curricula. There is no "consensus of public knowledge" on the nature of curriculum evaluation which warrants methodological formalizations about its character or provides the substance of such formalizations.

The curriculum research reviewed here is limited to literature which appears in the professional journals, and which attempts to evaluate curricula. This represents only a small part of the total writings on the subject. The narrative writing describing curricula and discussing theoretical issues is mostly omitted, which simplifies the summary herein presented because evaluation has not dominated the curriculum scene by any means. The subset of evaluation studies is much smaller than the set of curriculum development programs. In general, evaluations have not led to many encouraging findings, although evaluations—because of the complexity of the process—generally lack sufficient scope, so that an absence of positive findings is not surprising. Westbury (1969, p.245) summarizes the problem of matching evaluation schemes to curriculum objectives:

> Two separate, though interrelated analytical problems must be faced: curriculum must be conceptualized in such a way that it no longer carries the connotation that it is a unitary notion, often a treatment; evaluation must be seen in ways that permit the development of sets of methods and criteria

so reasoned judgments, appropriate to all senses of curriculum, become possible. Curriculum evaluation theorists must attempt to formalize these criteria and methods so they can prescribe rules for the application of criteria to the full range of concrete curricular issues.

No current theoretical prescription for curricular evaluation approaches these goals, although parts of the problem have been acknowledged by some writers.

Curriculum development programs in science and mathematics have been developed and evaluated, at least in some aspects. Some of these are reviewed by Romberg (1969), Smith (1969), Welch (1969), and Westbury (1970). Evaluation studies of curricula developed by the Physical Science Study Committee (PSSC), Biological Science Curriculum Study (BSCS), Chemical Education Materials Study (CHEM), and School Mathematics Study Group (SMSG) are inconsistent in their findings. Oftentimes, differences between these curricula and conventional ones are small, and sometimes results favor the conventional method. Some interactions are noted between student ability and measure of learning for different curricula; i.e., low ability students may do better in the conventional curriculum in terms of one measure of learning, while they do poorer in a new curriculum. All learning measures do not disclose this interaction, and on some of them the new curriculum is better (Welch, 1969, p. 439).

Westbury (1970, p. 250) summarizes a study by Heron (1969) that showed how a teacher's misunderstanding of a program might affect the program's success or failure. Heron made no attempt to evaluate curricula in terms of output measures. Rather, the study explored three evaluative questions related to CHEM, PSSC, BSCS curricula:

> (1) To what extent is the "enquiry" objective of these programs actually embodied in the materials produced? (2) How do the teachers through whom the materials filter perceive this objective and do they understand "enquiry" well enough to operationalize any conception of what it might mean in their classrooms? and (3) How does this objective compare to the explicit and implicit goal teachers set in their classrooms?

Westbury summarizes the findings:

> The results of his application were disappointing. Despite the claims of the developers for their materials, they were found to present little more than a "somewhat sophisticated" version of a "less competent" view of method. The teachers who had been attending workshops on the new materials were found to have almost no conception of what might be meant by a claim to teach the "nature of scientific enquiry."

The innovative science curricula such as those discussed above place heavy emphasis on the role of inquiry or learning by discovery, an emphasis that Ausebel (1965, p. 259) has severely criticized:

> Much of this "heuristics of discovery" orientation to the teaching of science is implied by the view that the principal objectives of science instruction are the acquisition of general inquiry skills, appropriate attitudes about science, and training in the operations of discovery. Implicit or explicit in this approach is the belief that the particular choice of subject matter chosen to implement these goals is a matter of indifference (as long as it is suitable for the operations of inquiry), or that somehow in the course of performing a series of unrelated experiments in depth, the learner acquires all of the really important subject matter he needs to know.

Later in this section theories of instructional organization are discussed (including Ausebel's). These approaches emphasize the importance of instructional structure in acquiring knowledge. It is not surprising then that Ausebel should conclude that incidental learning as a by-product of discovery cannot compare with a graded and systematically organized approach.

The idea of learning by discovery has become a popular one throughout education, particularly among those who are calling for reforms in classroom teaching. The complex issues involved in this concept are the topic of an excellent book edited by Shulman and Keisler (1966). The book emphasizes that learning by discovery does not mean laissez-faire education. The difference is in the way control is exerted, not in the lack of it. In general, learning by discovery has not been proven to have a great advantage over conventional methods. Cronbach (1966) points out that research is needed which attempts to determine what advantages learning by discovery offers, and under what conditions its benefits are accrued.

Although curriculum development is far from being on firm ground, and in spite of a general lack of evaluation, some progress is being made. The current status of curriculum development and evaluation in terms of its accomplishment and shortcomings are seen in the following quotations:

> In brief summary, during the past decade significant progress has been made in the precise definition of curricular objectives, in the analysis of ends/means relationships, and in the effective ordering of stimuli for learning. Substantial progress has been made in extending both the understanding of the evaluative process and the use of evaluative data in diagnosing the possible causes of discrepancies between curricular expectancies and curricular accomplishments. In the realm of explaining curricular realities, however, we appear to know little more in 1969 than we knew in 1960. Curricular theory with exploratory and predictive power is virtually nonexistent. Goodland (1969, p. 374).

> Research during the period of this review shows a desirable tendency toward a broader spectrum of concern, but still lacking are systematic longitudinal studies showing the impact of varied methods and materials on student attitudes, understanding, performance and motivation. Current research seems to be mainly discipline-centered rather than pupil- or learning-

centered, and the ends of education appear to be too often subordinated to transitory fashions in educational haberdashery. Smith (1969, p. 409)

One conclusion seems obvious. Only at centers where there has been a concentrated effort to investigate many facets of a course or teaching method by a group of researchers does one find any discernible evidence of advancement. Welch (1969, p. 441).

Theory must inform the deliberation that is evaluation but at the same time it must grow from deliberation. The problem implicit in this assertion is mapped by the requirement that curriculum and evaluation workers find a theoretical structure that permits them to embrace the particular and concrete with seriousness before they attempt theoretical speculation of any kind. We are far from this at the moment. Westbury (1970, p. 257)

Rosenshine (1970a) indicates that a central problem in evaluation is determining the actual teaching practice that takes place within any given curriculum. Because teachers vary widely in their skills, attitudes, beliefs, and dispositions, they do not all do the same thing given the same curriculum. Simply producing a curriculum does nothing in terms of its implementation, and evaluations of different curricula are generally useless without data on classroom transactions.

Rosenshine (1970a, p. 296) points out the almost complete lack of evaluative studies that include data on classroom transactions. In summarizing the shortcomings of evaluation of curricula, he states:

Currently, three major needs are: greater specification of the teaching strategies to be used with instructional materials, improved observational instruments that attend to the context of the interactions and describe classroom interactions in more appropriate units than frequency counts, and more research into the relationship between classroom events and student outcome measures.

Some progress is being made in defining classroom transaction and relating it to student outcome. Some studies that relate specifically to the teacher's mode of presentation were discussed previously; however, as yet there is little demonstrable evidence for accepting any particular curriculum as being better than another. This is a gross generalization and perhaps does not do credit to some programs. Of course, some curriculums are undoubtedly better than others and "everyone knows it." Unfortunately, demonstrating curriculum effectiveness is extremely difficult.

Instructional method studies have failed for essentially the same reason as curriculum studies: a lack of classroom transaction data. Reported studies find no consistent indication for the superiority of any instruction method. For example, research on discussion versus lecture has a long history, but as Stephens (1967, p. 81) concludes: "It has been found in summary after summary that no distinction between the two methods can be found."

Studies of instructional method rarely control for student or teacher character-

istics, and it is entirely possible that one method may be superior to another for some students and with some teachers. It is unreasonable to assume, for example, that all teachers are equally effective using the discussion method, or that because one is effective using the discussion method, he will also be effective using the lecture approach. Before instructional methods can be evaluated, certain student and teacher characteristics must be defined, and data must be provided on the transaction between them.

Television and Programmed Instruction

Teaching machines, programmed instruction,[10] and in general more technologically oriented aspects of instruction are described now. Many reviews of teaching technology have been written, but only the major ones are mentioned. Saettler (1968) provides a detailed and lengthy history, as well as a critical and summary review of research. W. H. Allen (1971) gives a brief overview of history and research, which includes comments on general research shortcomings. Chu and Schramm (1967) provide a lengthy evaluation and review summarizing research on televised learning. A number of other reviews of specific areas are cited in the following pages.

The early and intense interest in television learning led to a frenzy of development, with very little attempt at controlled research. The usual promotion claims were made for the success of these programs. Subsequent research did not support the claims, although as Chu and Schramm (1967, p. 176) point out

> In a sense, instructional television is more complex than the research that deals with it. Complex behavior has baffled learning theorists for years. A number of variables are clearly at work determining what a given individual learns from the television. In many cases these variables interact, and the total must be a great deal more complex than can be represented by the one variable experiments that typically make up the research literature, no matter how clean and skillful they are.

Within these methodological limitations, and after hundreds of studies, it appears that televised learning is about as effective as conventional classroom learning, and a case cannot be made for the superiority of either. Effective television teaching grows out of the application of sound teaching methods, such as simplicity, material organization, practice, etc., and apparently not from any special feature of the mode of presentation. Television, of course, reaches a larger audience and augments conventional methods. Further research must determine under what conditions television learning takes place and what the specific factors are in television presentation responsible for learning. However, the same comment holds for conventional teaching.

[10] Programmed instruction refers to the detailed sequencing of instructional tasks and is planned to produce continuous activity on the learner's part, with immediate feedback concerning the correctness of his response (see Corey, 1967).

The most direct application of learning principles has been in programmed instruction. This literature is reviewed in many places and is commented on in almost every review of educational research. Emphasis on programmed instruction surged a little over a decade ago, but interest has waned considerably over the past five years (Corey, 1967). Based on the operant conditioning paradigm of Skinner (e.g., Skinner, 1968), and following from his success in conditioning animal behavior, the same techniques were applied to human learning. Despite the early bloom and rapid spread of programmed instruction based on the Skinnerian method, later evaluations of the effectiveness of programmed instruction are not highly positive. The behavioristic learning approach of Skinner and his followers was criticized early in its development on the grounds that they had derived their teaching practices from work with animals, making the programmed instructions void of highly meaningful structure with too much concentration on rote-type material. Some criticisms by Pressey (1963), and Thelen (1963a,b) relate specifically to programmed instruction. Of course, the Skinnerian stimulus-response approach drew instant fire from their old antagonists, the Gestalt psychologists, who insisted on a field[11] approach with emphasis on meaningful units instead of fragmented, serially presented (and rote-learned) programs.

Theoretical issues aside, programmed instruction has not proven the success it was thought to be in its early days (Gotkin and McSweeney, 1967; Saettler, 1968; Allen, 1971). It is about as effective as conventional programs when student achievement is used as the criterion, but its superiority has not been affirmed. Few, if any, of the claims that have been made for the efficiency of teaching machines have been proven, despite untested claims made by teaching machine manufacturers, and the fervor of their sales promotion (Saettler, 1968, p. 269). An early claim for programmed instruction was that by properly sequencing material in small steps, the dull student would be able to perform better and, some claimed, even as well as the bright ones. In their review, however, Cronbach and Snow (1969) could find no clear evidence to support these claims.

In summary, programmed instruction has not been proven superior to conventional classroom methods, and this probably explains the recent decline in research on the topic. But a number of positive outcomes have grown out of the interest in programmed instruction and teaching machines. A recent book on programmed instruction, edited by Lang (1967), has little to say about programmed instruction as it applied to teaching machines; rather, the book deals with the design, structuring and sequencing of learning material for any mode of presentation, including problems of curriculum design.

Allen (1971) points out that programmed instruction research has produced an interest in developing individualized instruction. Whereas early research and application focused on group instruction and one-way communication, the current interest is shifting to the unique characteristics of the individual student as a central issue in instruction design. The interest is turning, however slowly, to the study of the interaction between student, task, and material. Development and research on

[11] Very loosely a "field" refers to an individual's total environmental and behavioral complex in time.

individualized instruction is in process at several centers, and in general the results appear promising. Application of such techniques depends on factors including instructional content and possibly some learner characteristics. Because of the complexity of this field of research, and because there are apparently no recent reviews, this topic is not pursued further.

EXPERIMENTAL WORK IN INSTRUCTION

Organizing and making relevant to instruction the vast psychological research is an enormous and perhaps even an impossible task. Gagné and Rohwer (1969, p. 381) have stated the problem well:

> Remoteness of applicability to instruction, we note with some regret, characterizes many studies of human learning, retention, and transfer, appearing in the most prestigious of psychological journals. The findings of many studies of human learning presently cannot be applied directly to instructional design for two major reasons: (a) the conditions under which the learning is investigated, such as withholding knowledge of learning goals from the subject and the requiring of repetition of responses, are often unrepresentative of conditions under which most human learning occurs; and (b) the tasks set for the learner (e.g., the verbatim reproduction of verbal responses, the guessing of stimulus attributes chosen by the experimenter, among many others) appear to cover a range from the merely peculiar to the downright esoteric. This is not to imply that such studies do not further an understanding of the learning process. However, it would seem that extensive theory development centering upon learning tasks and learning conditions will be required before one will be able to apply such knowledge to the design of instruction for representative human tasks.

Much of the difference between learning experiments in the laboratory and in the classroom must lie in the influence (direct or indirect) of behaviorism, which is based on stimulus-response relationships and control through the manipulation of reinforcement. The inadequacy of this technique, even in simple animal learning, has been questioned repeatedly, and its application to human learning (particularly verbal) is considered by many to be grossly inadequate (e.g., Deese, 1969; Garrett and Fodor, 1968). There have always been severe critics of the behavioristic tradition in general. Recently, the psycholinguists, led by Chomsky (1959), have leveled some devastating criticisms, and the debate continues. Although other theoretical formulations exist, behaviorism dominates in learning and experimental psychology; the methodologies used in learning studies are almost exclusively of the behavioristic type. Some examples of widely used methods are summarized below.

The method of association learning typically presents pairs of stimuli (words, symbols, pictures, etc.) to the subject during the learning phase, and tests for learn-

ing by presenting him the first stimulus and testing for his recall of the second. A recognition measure of retention (or learning) may be used in which the subject selects the correct stimulus out of several presented to him. An even more primitive form (serial learning) simply presents stimuli in lists, and learning is measured by the degree of recall (or recognition) of list items. In analyzing human learning, hundreds of laboratory studies involving serial and association learning occur each year, but the value of studies of paired-associate learning for classroom-type learning has been repeatedly questioned, and it is generally concluded that their value is minimal. However, Rohwer et al. (1971) caution against this conclusion because substantial relationships have been reported between paired-associate and school learning.

Another frequently used method is that of *discrimination learning*, in which the subject learns to respond differently to different stimuli through the application of a reinforcer. Usually there are two stimuli and two responses. For example, the subject may be reinforced (with a reward or with feedback concerning the correctness of his response) for responding to one stimulus, and not reinforced for responding to another. Learning is measured in terms of the time or number of responses necessary for the subject to "learn" to respond only to the "correct" (the reinforced) stimulus. This method may make use of an irrelevant stimulus (one present but one not necessarily attended to by the subject), and the subject is then tested for how well he "correctly" responds to this incidental stimulus (incidental learning).

At least two excellent reviews of instructional research are available in the last few years—by Anderson (1967) and by Gagné and Rohwer (1969). Both organize research around a few central issues, and both evaluate as well as summarize research as it relates to these issues.

Transfer of Learning

A central issue in learning theory, and a critical one in classroom learning, involves transfer or generalization of learning. Preschool and compensatory education programs have been disappointing because achievement gains have faded over time. This has led to an interest in how achievement in basic skills such as reading and math might generalize to future achievement and to concurrent achievement in other school subjects. Apparently, no attempts have been made to measure this generalization in the classroom directly; however, psychological research on generalization (mostly referred to as transfer) is vast. Gagné (1962) distinguishes two kinds of transfer. One transfers the learning of a specific task to performance on the same general class of tasks. He terms this *lateral transfer*, and it is the same as generalization, which operates whenever two learning problems require common rules for solution, or both depend on some common stimulus and/or response sequences.

Lateral transfer is becoming a less popular research topic (Gagné and Rohwer, 1969), and recent studies are apparently finding nothing new. Much research on lateral transfer has centered on learning general rules, in which case verbalizing the

rule is better than not, and using many examples of the rule in the learning phase helps to promote transfer.

A second kind of transfer, termed *vertical*, operates when the learning of a specific task facilitates the learning of another. For example, training in stimulus coding transfers to paired-associate learning; i.e., subjects trained in coding learn faster (stimulus coding entails a translation of meaningless symbols into meaningful ones by association—a mnemonic device). In such transfer, stimulus coding is a subordinate skill to paired-associate learning; however, it is not necessary to, or a part of, the learning task. This is the kind of transfer that Gagné and others consider in studies of hierarchical organization.

Vertical transfer studies carry a number of important implications for instruction design. Gagné (1962) first outlined the notion of hierarchical organization, as described earlier. Theory predicts that in learning a subject, students cannot "pass" a post-test unless they have also "passed" tests for skills lower in the hierarchy of knowledge. A number of studies designed along these lines support Gagné's theory. In a more recent review, Gagné and Rohwer (1969) state that "Studies of transfer of prior learning are frequently consistent with this hypothesis, although few are confirming in a crucial sense."

Asubel (1963) developed a theory of hierarchical organization of meaningful verbal material. The hierarchy begins at the bottom with detailed and specific bits of knowledge and builds to a level containing the most abstract and general concepts. The learning of new material can be facilitated by using "advance organizers," which help the learner integrate new material into his existing cognitive structures. Such organizers are highly generalized statements or questions that the subject reads prior to studying new material in order to prepare him for new material in terms of what he already knows, or to outline and brief the material. In addition to experimental support cited in the original article, several other studies also find supportive evidence for the theory (Allen, 1970; Grotelueschen and Sjorgren, 1968; Merrill, Barton and Wood, 1970; and Merrill and Stolurow, 1966).

Vertical transfer has been studied under a number of other theories and experimental disciplines including rule learning, concept learning and attainment (see discussion of Piaget's work in Section V), verbal learning and problem solving. Many results clearly indicate the importance of the sequence of tasks on instruction effectiveness. These results appear to have more direct bearing on classroom learning than any others we have reviewed, although much more needs to be known.

A topic closely related to transfer involves a technique called "fading" or "vanishing," in which one stimulus is faded out and slowly replaced by another. Anderson (1967) reports that research in this area may have practical value for teaching children who cannot understand or hear verbal instructions. The students are able to learn to make the correct response to the new stimulus without trial and error behavior. A recent study by Karraker and Doke (1970) found the fading technique to be superior for errorless learning by kindergarten children in discriminating between letters b and d; however, Samuels (1970) summarizes reading research using the fading technique and finds contradictory results. In fading, a picture and a word are shown together, and the picture is gradually faded out. It appears that

the attention shift from the picture to the word does not always take place. In view of the contradictory evidence and the limitations of this technique, it appears to have little classroom utility.

Reinforcement and Feedback

Reinforcement is a central concept in almost all learning formulations, and many learning theorists and experimentalists insist that learning cannot occur without reinforcement. Without a clearly defined external reinforcer, these theorists assume that reinforcement is provided by the subject and is internal. For example, a subject may be reinforced with some tangible reward for reading, or he may read because he finds it personally rewarding. The latter is considered to be a case of intrinsic reinforcement. Other learning is said to take place as a result of the operation of social reinforcers or broadly generalized extrinsic ones. The importance of reinforcers to learning has been realized in the laboratory through the strict use of an operational definition (e.g., if a stimulus presented immediately after a response leads to an increase in the response rate, it is a reinforcer). It is frequently argued that using this operationally defined reinforcer concept in complex learning is at best unproductive. The stimulus properties of the reinforcer are not known, nor is the response that is to be reinforced well-defined. The reinforcement concept, when carried to its limits, becomes virtually tautological, and therefore of little practical value in educational research.

A general term some psychologists use to indicate an information processing and volitional aspect of complex learning is "feedback," which may be used to denote either the reinforcing event, the subject interest in and use of the event, or both. Thus, obtaining a penny (or candy, etc.) reward for the correct response in a discrimination learning task may be thought of as providing feedback about the correctness of response and defines how the subject can obtain further reward. Many theorists feel that providing knowledge of results to the learner (feedback) reinforces the desire (drive, for example) to learn and that the reinforcing event is primarily intrinsic, although under partial control of the external event (feedback).

Although studies of reinforcement factors have dominated much of the psychological literature on learning, it appears that very few results have any real value in determining classroom learning. Using a term like "feedback" has not clarified the issue. Certainly, reinforcement and feedback are not consistent factors in conditions of learning, and Gagné and Rohwer note that (p. 401):

> A characteristic of recent research is that it reveals clearly the highly variable nature of feedback effects. Moreover, the research indicates that the sources of this variance are to be found in learner characteristics, type of feedback, timing of feedback, direction of feedback, and type of task.

Attention Factors in Learning

For learning to occur, the learner must pay attention to the appropriate stimuli, and attention factors have played a central role in learning experiments. There is

a well established body of research to indicate that stimulus novelty promotes learning and helps to maintain attention. In human learning, it is found that guessing and delayed feedback lead to better learning than no guessing and immediate feedback. In general, factors that increase the uncertainty of a stimulus complex lead to heightened curiosity and/or increased attention. In reading material, retention is improved by inserting questions throughout the text. These results are generally indicative of increased attention and inspection time.

One of the more easily manipulated factors in instruction is the mode of presenting the learning material. In summarizing the research on stimulus presentation, Gagné and Rohwer (1969, p. 394) state:

> Considerable evidence has now been amassed indicating that when there is a choice of method for presenting equivalent information, the following results prevail: pictorial materials are superior to verbal; concrete verbal materials are preferable to abstract verbal; and grammatically structured are better than unstructured materials. In contrast, the conditions that might dictate choices among various available modes of presenting stimuli are almost entirely undetermined thus far. Finally, stimulus context appears to be one of the most potent of the variables determining the effects of materials presented, although tasks other than traditional laboratory ones remain to be investigated.

Research that finds pictures superior to words is mostly based on the paired-associate method, which typically requires the subject to learn lists of paired words, paired pictures, or a word paired with a picture. Although results favor the picture presentation, the relative effectiveness of either mode appears to depend on many factors including student characteristics and age, and task characteristics. On the basis of classroom studies, however, Samuels (1970) finds that pictures negatively affect learning to read, especially for the poorer students. He interprets pictures as distracting stimuli that produce attention shifts. This is consistent with other findings about how distracting stimuli affect learning by poor students. Samuels' studies involved young children learning to read, while most of the studies using the paired associate method used older subjects. The age difference may account for the disparate results obtained from the two methods.

Retention of Learned Material

Once material has been learned, how much of it will be retained? Studies of retention and forgetting are as old as the study of learning, and one of the principal measures of learning has always been the amount of retention. Gagné and Rohwer (1969, p. 401) give an excellent review of the research, the principal findings, and the basic issues involved. Unfortunately, like much of the research reviewed here, the data on retention are mostly based on the paired-associate method, which makes generalization to the classroom hazardous.

Earlier studies which seemed to demonstrate better retention for free recall

compared to recognition learning have since been shown to be a function of the degree of original learning, rather than the method of learning. A number of studies agree that when control is introduced for the degree of original learning, retention is approximately the same for all learning methods (within the limitations of paired associate learning). Even the material's degree of meaningfulness does not affect retention when the degree of learning is controlled. Of course, "meaningfulness" here is used strictly in the framework of paired-associate learning, where meaningfulness refers to the use of words instead of nonsense syllables, or the use of grammatically correct sentences compared to random word orders. This does not seem to be closely related to what educators generally mean when they talk about meaningful material.

Other factors affecting retention have been isolated. The effect of retroactive inhibition is well known. This occurs when a learning task inserted between the learning of an original task and the retention measure causes the student to forget the original material. It has also been found that elaborating (talking, etc.) on the stimuli in the learning phase promotes retention.

SUMMARY

Instructional method studies on classroom and curriculum design have produced no clear and consistent results. The problem, again, is basically one of evaluation and a lack of adequate data. Television learning, teaching machines and programmed instruction appear to have no general superiority over conventional methods, although they can reach more students. Psychological studies of the factors affecting instruction use tasks generally different from classroom learning tasks, and as a result they tend to have limited value for determining instructional methods. A promising area of research concerns transfer of learning, especially organizing instructional material. Apparently, instructional material can be organized in a hierarchy, although the rationale and basis for the organization is not clear. More research is needed on almost every factor being studied. In addition, data are badly needed to bridge the gap between laboratory and classroom. There is some discussion and some evidence about the importance of interaction among the individual student, the instructional method, and the type of learning task, but this area has hardly been touched.

IMPLICATIONS

This section implies that a data system must be able to record information on the individual student's progress. Normative evaluation is important to indicate the overall success of the student and the program, but the greatest importance of achievement data seems to be their value for remedial evaluation. This is also true

if criterion referenced tests are used. Specific instructional methods will produce requirements for certain kinds of data storage and retrieval, but they cannot be specified in the general case. The methods discussed above require that the student be identified and that data are accessible during, as well as after, a course of instruction.

V. STUDENT CHARACTERISTICS

INTRODUCTION

This section describes how a general failure to match student characteristics to specific educational programs is a major reason for the lack of positive findings in educational research and the consequent lack of success in defining factors that substantially affect educational outcomes. Little has been done in developing specific educational programs to fit individual characteristics. *A priori,* however, it seems reasonable to believe that students respond differently to different kinds of classroom and instructional methods, and to different types of teachers. As reasonable as this hypothesis may sound, there is little research to support it, although some notable exceptions are pointed out below.

Undoubtedly many social reasons exist for bypassing individual student differences as a major part of research; we note reasons internal to psychology. Cronbach (1957) points out that psychology was split into two disciplines. One group (mostly psychometricians, and to some extent personality theorists) has been concerned with individual differences (differential psychology) and has paid little attention to developing a general behavior theory; another group (notably learning theorists and experimental psychologists) has attempted to develop behavior theories while ignoring individual differences. This split has been particularly damaging to education because learning theorists consequently have little to say that bears directly on classroom learning. Gagné (1967, p. 13) noted

> First the widespread inattention to individual differences seems to indicate the psychologists have been uniquely optimistic in their expectations for the generality of behavioral laws. In the pursuit of these laws, the assessment of ranges of generalization and of limiting conditions has been by-passed. If we recognize learning as a process of transition from an initial state to an arbitrary terminal state, then with respect to the individual differences problem, we should take a lesson from other natural sciences. We must recognize limitations in the applicability of a scientific law. It is through the specification of limiting conditions that our hypothesized or theoretically derived relationships obtain concreteness.

The following subsections discuss evidence for the importance of individual characteristics in determining educational outcomes. Some evidence reviewed is directly associated with classroom learning; however, most of it is less direct, originating in studies of personality, developmental psychology, and differential psychology. Studies in these categories seldom use conventional classroom learning as an outcome measure (dependent variable).

ABILITIES AND GENERAL INTELLIGENCE

The study of individual human abilities has long been an area of psychological research. The various theories and the literature generated by this effort are reviewed in many places (for example, Guilford, 1967; Cronbach and Snow, 1969; and Snow, 1971). The most generally accepted theories identify some kind of general ability (general intelligence) and a number of special abilities.

The relative influence of heredity and environment on the development of abilities is a topic of continued interest and heated debate. Some theories assume that abilities are genetically determined and unfold in the development process. Others maintain in varying degrees that abilities are learned and that heredity only places loosely defined boundaries on their development. Snow (1971) comments that "The bulk of the evidence seems to be against the unfolding hypothesis, but the alternative learning hypothesis remains largely untested."

The most recent upsurge of interest in genetic determinants of intellectual ability was prompted by the work of Jensen (1969), who describes the interaction of two broad categories of ability (Level I and II) and type of learning (associative and conceptual). Jensen's findings and his interpretation in terms of heredity are a matter of much controversy, and more research is needed before any firm conclusion can be made. In particular, the effect of "tuning"[12] on students low on tests of Level II ability must be investigated because there are subjects who have had little exposure to, or use for, conceptual thinking.

A recent, well-designed study by Rohwer et al. (1971) investigates several hypotheses derived from the Jensen model. Some results support the model, others conflict with it. The authors also present an alternative explanation that does not depend on differences in innate ability between populations. Part of the problem in verifying Jensen's model is that Level I and II tasks are not readily defined.

Although the relative contributions of heredity and environment are unknown, evidence confirms differences in general cognitive performance between ethnic groups. Stodolsky and Lesser (1967) review the evidence on this subject and describe their own carefully controlled study in which they find highly significant differences

[12] Some students have little or no practice in the use of mediation or looking for general principles in problem solving. Thus, they do poorly in abstract or conceptual problem solving when compared with children who come from an environment that encourages the use of mediation. Tuning is a pre-training to teach subjects the use of mediation. Differences between groups often disappear when tuning is employed.

in achievement patterns across four mental abilities (verbal, reasoning, numbers, and space) for various ethnic groups (Chinese, Jews, Negroes, and Puerto Ricans). Attainment level for each of the four abilities varied within an ethnic group, but ethnic groups differed in terms of which ability they attained best. Differences were also found for lower and middle class children within an ethnic group, and while the patterns were different for different ethnic groups, they were nearly identical for the two classes within an ethnic group. Thus, whatever factors operate to produce the differences in ethnic patterns of mental performance, also operate in both lower and middle classes. Stodolsky and Lesser point out that more research is necessary to determine the specific antecedents of the differential patterns of mental ability.

Some recent successful attempts to improve IQ scores of Negro ghetto children argues against a genetic explanation of the generally lower scores. Through working with parents, some recent attempts to modify IQ in preschool children show promise as do some programs that focus on language learning (see Elkind and Sameroff, 1970, for a review of these studies). Two recent promising programs which began with preschool children were a University of Illinois project (Engelmann, 1970), and the Ypsilanti-Carnegie Project (Lambie and Weikart, 1970). The Illinois programs especially demonstrated substantial gains in IQ scores and school achievement. But past studies have shown that over time IQ gains resulting from special programs decline, so that one needs to know the longitudinal effects before making a final evaluation of these programs.

The above studies are examples of success in identifying special abilities. The important question is to determine how these abilities affect educational outcomes. Studies investigating the effect of special abilities on learning have been summarized and evaluated in a number of places (e.g., Ferguson, 1965; Fleishman and Bartlett, 1969). Cronbach and Snow (1969), however, find serious methodological flaws in much of this research and conclude that there is little clear evidence for assuming an interaction between special abilities and learning. This is not meant to imply that specific abilities do not affect education outcomes, but rather that their utility for differentiating success among particular teaching methods has not been adequately demonstrated.

Whether or not general intelligence (or general ability) is related to learning is a controversial matter. Evidence from factor analytic studies indicating that intelligence is not a unitary ability, and low correlations from studies of IQ and learning and between several learning tasks, led Fleishmen and Bartlett (1969) to favor an interpretation that does not define intelligence as the ability to learn. Cronbach and Snow take issue with this point of view; after reviewing and reanalyzing existing data, they conclude that general intelligence is consistently and substantially correlated with learning. Much of the confusion, according to these authors, has arisen because many studies of the relationship between intelligence and learning use laboratory tasks that do not allow general intelligence to have much effect. In addition, most of the support for special abilities comes from the factor analytic approach that dominated American research on abilities for several decades. This technique tends to overdifferentiate because even slight correlations sometimes produce new factors and, in the process, a general intelligence factor tends to be

submerged. British researchers have used a hierarchial model of abilities (e.g., Vernon, 1965). The views of Cronbach and Snow are more consistent with the British approach.

Along with correlating general intelligence with degree of learning, Cronbach and Snow (1969) report evidence of significant and substantial interactions between intelligence and instructional method (aptitude-treatment interaction). In other words, instructional methods and learning tasks can be found that have different effects based on a student's general ability. For example, given instructional methods A and B, an interaction effect means that if high ability students do relatively well under treatment A, they do relatively poorly under B. Conversely, low ability students do relatively well under B and poorly under A. If groups of students given treatments A and B are equally mixed in regard to ability, then no difference will be found between the average performance under the two treatments. This is considered the reason for much of the failure to find positive effects due to instructional innovation. The kinds of education treatments that will produce such interactions with general ability are not well understood, but some possibilities can be brought out in the following pages.

In view of an interaction between educational method and student intelligence, then, to maximize achievement, students should receive different types of instruction (at least in some topics) on the basis of intelligence. But classroom grouping by intelligence or any other ability has a long history of failure in promoting any difference in achievement outcome. Thelen (1967) reviews the extensive research on grouping and summarizes the findings of the international conference on grouping at the UNESCO Institute of Education in Hamburg in 1964. Results clearly indicate that heterogeneous groups do about as well as homogeneous groups. The reason for this seems obvious. Grouping, on any basis, by itself, cannot be expected to produce improvement. What is needed is differential instruction treatment of the separate groups as Thelen points out (1967, p. 188).

> In other words, special grouping makes sense only when the teacher has a clear and accurate idea of what to do with the special group. From this standpoint, the chief difficulty with homogeneous ability grouping is that the guesses about how to deal with the group are often wrong. Thus, we find teachers who think "bright" children "ought" to be more self-directing, more interested in the subject, more creative, or more eager to have a continuous, heavy load of work. By and large, however accurate these guesses may be with regard to impressions of bright adults who are successful in the adult world, the guesses are mostly not true—and certainly not necessarily true—as applied to most bright children under usual school conditions.

STUDENT CHARACTERISTICS AND PROGRAMMED INSTRUCTION

In the last decade an interest has developed in programmed instruction and the application of what is sometimes referred to as principles of learning theory. This

interest derived almost entirely from the psychological field of learning; as mentioned earlier, the discipline was not oriented toward accounting for individual differences. Therefore, most of this research, especially that on programmed instruction, has not focused on (or even considered) individual characteristics. Thus, most instructional method research was reviewed above. Here we summarize the findings of studies that have atttempted to investigate response to programmed instruction as a function of student characteristics.

Cronbach and Snow (1969) point out a study by Burton and Goldberg (1962) that is exceptional in its sophistication and leads to an interesting hypothesis requiring further investigation. Their essential finding was an interaction between treatment (type of feedback) and student aptitude (verbal reasoning), but the interaction reversed, depending on the difficulty of the learning task. This is particularly important because it indicates that higher-order interactions exist, as well as interactions between ability and task difficulty.

Another excellent study (according to Cronbach and Snow) that indicates not only simple interactions but higher ones as well, is that of Maier and Jacobs (1964), in which some classes in Spanish had programmed instruction (PI) only, some had PI plus live instruction, and others had live instruction only. In addition, students were tested for general intelligence, Spanish language ability, and attitudes about Spanish. Results indicate that a favorable attitude toward Spanish was associated with PI plus live instruction for high intelligence students, and with PI only or teacher only instruction in low IQ classes. Second, low ability students tended to favor PI while high ability students tended to favor live instruction. Perhaps the most significant finding was that some teachers got better results under one set of techniques and student characteristics than under others. It appears that high IQ students do better under PI plus teacher when the teacher favors the innovative method. We return to this topic later.

Although far from conclusive, evidence supports the notion that students with low aptitude (low general intelligence) may respond differently to some programmed features compared to students with high aptitude. Well-structured programs may be more effective for duller individuals, and perhaps brighter students respond better than dull ones on scrambled presentation. In general, however, support for an interaction between programmed instruction and student aptitude is meager.

STUDENT CHARACTERISTICS AND MEANINGFULNESS

The issue of meaningful versus rote learning has a long tradition, and introductory psychology texts will usually say that meaningful material is more easily learned. Rote learning is generally considered to require less ability, and one is led to expect an interaction between meaningfulness and ability.

Cronbach and Snow (1969) surveyed the research on the effect of meaningfulness of instruction and its interaction with student aptitude, noting some evidence

of an interaction. It is not clear, however, what factors actually allow one type of student to gain more from meaningful instruction than others. Tuning is seldom used so that students who have little or no experience with meaningful material are not on a par with students who have. Cronbach and Snow (1969) comment on a large-scale well-designed study by Brownell and Moser (1949) which investigated meaningful versus mechanical instruction in subtraction. They state (p. 108):

> In half the schools, subtraction was rationalized for the children; a major effort was made to explain why certain steps were performed in (e.g.) borrowing. But third graders in some of the schools seemed unable to profit from these explanations. The authors tell us that where instruction had been rote in the two preceding grades the whole concept of explanation in arithmetic was strange to these pupils, and they could not incorporate the meanings offered. The children, then, had developed a positive inaptitude for meaningful instruction, whereas other children had been led to the point where they could profit from explanation. Now this is important first in undermining the concept that aptitude or readiness is simply a matter of intellectual maturity. Secondly, it sharply challenges such a concept as Jensen's regarding a native incapacity. Third, it destroys any lingering attempt to define "one best way" of instruction. Fourth, it urges us in the direction of trying to help the pupil who does not use meaningful instruction effectively by combining techniques that will move his skills forward without relying on comprehension, with techniques that will advance his ability to comprehend. We are in no position to write off these third graders as noncomprehenders—but we do not anticipate that simply tuning will bring them to the level of mathematical reasoning.

A series of articles on the use of advance organizers in learning meaningful verbal materials (previously reviewed) culminated in a recent study by Allen (1970) describing evidence of aptitude-treatment interaction. Advance organizers are highly generalized statements read prior to learning new material. These statements facilitate learning by allowing the student to relate the new material to his existing cognitive structure. Results indicate that the advance organizers facilitate learning (measured by delayed retention) in higher ability students, but not in lower ability ones. This may indicate that students of lower ability do not have the cognitive structure necessary to make use of the advance organizers. This study raises a number of interesting questions that need further exploration.

CONCEPT ATTAINMENT

One area of major interest to psychologists, particularly in the field of child development is concept attainment and cognitive development, in which studies

attempt to determine the sequence of concepts as the individual attains them or relates them to each other. Several different theoretical explanations and experimental approaches to the study of concepts have been taken, and Gagné (1968) presents them in capsule form.

Learning theorists who belong to the associationistic school consider concept attainment as mostly a matter of learning. Others believe concept attainment depends on maturation and biological readiness. The most popular theory currently is that of Piaget, who focuses on the organism's existing cognitive structure in terms of its adaptation to its environment. Changes in adaptation are related to modifications in the cognitive structure. Gagné (1968) proposed a model based on cumulative learning effects (of which association is a small part) within limitations imposed by maturation. These and other models differ markedly in terms of the importance assigned to learning.

Concept development theories are directly relevant to education for they define the factors upon which levels of learning depend. If concept attainment is largely a matter of maturation and readiness, or level of cognitive structure, then the student should not be exposed to a task for which he had not developed adequate concepts. If concept attainment depends upon prior cumulative learning, however, then instruction must utilize only the prior learning that has occurred and sequence tasks in a hierarchy according to their contribution to other learning tasks. Gagné's theory of hierarchial organization rivals Piaget's ideas, although confirmatory evidence is still mostly lacking.

Regardless of which concept attainment theory proves most fruitful, differences do exist at a given time between students, and over time for a given student. These results have wide implications for instruction design and the time at which a student is exposed to specific instruction.

PERSONALITY DIFFERENCES

No field within psychology is more concerned with individual differences than the study of personality. No other discipline has a controversy as great, empirical findings less definite, and proliferation of theory as abundant. Reviews of this complex area are found each year in the *Annual Review of Psychology,* presenting several perspectives including the behavioristic approach (Sarason and Smith, 1971), the psychometric (Wiggins, 1968), the clinical (Klein, Barr, and Wolitzky, 1967), and others. Yet there is little that one can apply directly to education at this time, and methods for assessing personality traits are far from perfected, as noted by Sarason and Smith (1971, p. 397); "The pitfalls involved in attempting to assess significant personality attributes are many and varied, and the 'true score' of an individual's standing on a given dimension is as elusive as the Holy Grail." In spite of these pessimistic comments, some general results from personality studies have indirect implications for education.

A growing conviction and supporting evidence indicates that definitive person-

ality differences exist between the high and the low achiever. In reviewing the subject, Klein, Barr, and Wolitzky (1967, p. 534) summarize:

> High achievers show strong internalization of values, indicated by responsibility and socialization. They also have high achievement motivation, in regard to both independent and conforming spheres. They are, however, low on social desirability (need to make a good impression for its own sake) and lack flexibility, apparently preferring order and stability. The negative loading for flexibility appears in an equation developed on the Italian sample as well, as will be important when we come to consider what these studies reveal about the nature of the criterion itself. As Gough and Fink (1964, p. 380) point out, the pattern of the achiever "is not a pattern of creativity or innovation, but rather that of constructive adaptation to a world in which one's circumstances are modest and one's destiny limited."

Cronbach and Snow (1969) also discuss a study that shows an interaction between degree of meaningfulness of instruction and "overachievers" versus "underachievers." The "overachievers" showed better performance on the less meaningful material while the reverse was true and vice versa for the "underachievers."

The concept of anxiety is one of the cornerstones of personality theory, and has also become a major factor in learning studies. Adelson (1969, p. 231) began a review of the topic with this statement: "Anxiety was the most popular single topic in personality this year." And later (p. 233):

> After all these years, and after literally hundreds of studies of anxiety, there is still no general agreement as to what the commonly used scales are in fact measuring, whether it is drive level, maladjustment, affect, degree of defensiveness, or several of these in some interaction.

In the latest review, Sarason and Smith (1971) quote suggestions that much of the confusion results from a failure to distinguish between anxiety as a stable personality trait, and anxiety as a temporary emotional state.

Despite the confusion and ambiguity about anxiety, a few promising suggestions are possible. Many studies indicate an interaction between anxiety and intelligence on cognitive performance, such that anxiety enhances the performance of low ability students and deteriorates the performance of high ability ones. Cronbach and Snow (1969) describe an apparent interaction between personality and instruction. It appeared that structured instruction (as opposed to unstructured) was better for high-anxious, high-compulsive children. For the child who was neither anxious nor compulsive, both instructional methods were about the same. Cronbach and Snow point out that flaws in the design of the experiment make it dangerous to generalize, and it is possible that in some schools and for some students the unstructured method would achieve better results.

Student attitude and motivation are undoubtedly major determinants of achievement. In applied research, much work along these lines has attempted to change the student's attitude about education or to increase his motivation. Another

line of research, mostly done in the laboratory, has attempted to measure attitude and motivation and to relate outcome to them. Some studies have investigated the relation of motivation level to teaching technique and classroom structure.

A particular aspect of motivation that has received much attention is the achievement motivation, referred to as need-achievement. It appears that achievement motivation is a particularly persistent personality characteristic (Ryder, 1967) and one that is more related to cognitive maturation and innate ability than to early experiences or child rearing practices (Heckhausen, 1967). Other findings (reviewed in Dahlstrom, 1970; Flavel and Hill, 1969; and Hartup and Yonas, 1971) indicate that achievement motivation in young children has different antecedents than it does in adolescents. Adolescent and later achievement motivation is more related to parental and social rewards and punishments, whereas at a younger age it seems more related to an assertion of autonomy.

Cronbach and Snow (1969) review the literature on motivation as related to student-treatment interaction. Theory predicts interaction between need-achievement and education treatment, but attempts to demonstrate it experimentally have not been overly successful. Interactions are sometimes reported but they are small. The tasks used in most studies make it difficult to extrapolate to classroom learning. In addition, many of the studies are done with college students, and as pointed out above, there are indications of differential antecedents, depending on age.

The increased national interest in academic achievement (particularly reading and math in the early grades) has caused a certain amount of alarm concerning the possible negligence of other factors in student growth. The focus on achievement and the start of accountability system implementation to monitor and enhance certain cognitive skills introduce the risk of stifling noncognitive growth. Emphasis on rote learning (and it is generally agreed that most compensatory and achievement oriented programs emphasize rote learning) occurs at the expense of creative development. It is a popular lament among individuals identified as creative that formal education in many respects is a liability to creativity. While these self reports may not be particularly reliable, they should not be ignored. Research on creativity tends to support such notions, although studies of creativity are not highly definitive. In reviewing the research on creativity, Klein, Barr, and Wolitzky (1967, p. 536) note:

> Psychologists use widely different criteria in studies purporting to deal with creativity, ranging from the careers of eminent people (which are obviously worthy of consideration), to the idea of creativity in interpersonal relations (which makes one wonder whether this is really "creativity"), down to measures of sales productivity and customer service (which can cheerfully be ignored). Furthermore, even when outstanding achievement is the criterion, it usually does not include what most informed nonpsychologists consider to be creativity, that is, humanistic and artistic creativity.

Reporting on a study of creativity in children, Hartup and Yonas (1971, p. 377) report:

> . . . [there is] no clear support for the use of either tests or gamelike contexts in assessing creativity. Scores depend on the task, the measure of creativity, the anxiety level of the subject, and sex.

In summarizing recent creativity studies, Dahlstrom (1970, p. 34) states:

> At the present time, therefore, available evidence suggests that the creativity process involves a variety of enhancing variables: interest, involvement, sensitivity, and self-confidence; and a variety of inhibiting variables: fears, self-doubts, and disabling sets and misperceptions acting jointly to determine the degree of expression of whatever the level of skill and proficiecy of the individual for that situational demand will permit.

Dellas and Gaier (1970) provide an extensive review and penetrating analysis of the problems, issues and results in studies aimed at identifying creativity. The research on creativity is marked by a glaring deficit of replicative and follow-up studies, but despite these deficiencies, the authors are able to conclude (p. 67):

> Despite differences in age, cultural background, area of operation or eminence, a particular constellation of psychological traits emerges consistently in the creative individual, and forms a recognizable schema of the creative personality. This schema indicates that creative persons are distinguished more by interests, attitudes, and drives, than by intellectual abilities. Whether these characteristics are consequences or determinants of creativity or whether some are peripheral and of no value is moot. These questions remain insufficiently approached and elucidated.

Evidently, no one is in a position to write a formula defining creativity. It is equally apparent that in spite of many problems with the research, much is known about the characteristics of creativity. The creative person appears among other things to be independent in attitudes and social behavior and not very concerned about his impression on others; an education program mainly interested in behavioral conformity and standardized achievement has little of positive value to offer him. Accountability systems which at present can only focus on achievement in rote learning may further alienate the creative student, especially in the early school years.

EARLY DEVELOPMENT AND LEARNING

Psychologists, and especially psychoanalysts, have long stressed the importance of the very early years in the development of persistent behavior patterns. The time to affect cognitive and noncognitive development factors is during the preschool years. Kagen (1970, p. 9) writes:

> The idea of this suggestion rests on the assumption that a child's experience with his adult caretaker during the first 24 months of life are major determinants of the quality of life motivation, expectancy of success, and cognitive abilities during the school years.

He then reviews data which support this suggestion.

Support for the importance of early development comes from a wide variety of research examined every year in the *Annual Review of Psychology* under the heading of Development Psychology. Other support comes from the recent and growing interest in "critical periods" of development during infancy that determine life patterns. Most of this research has been conducted with animals, although there is supporting evidence from research and observations on humans.

The importance of early experience for education is the topic of a book edited by Denenberg (1970). The material is somewhat slanted toward the growing interest at a federal level in day-care centers, and toward the conviction that any really meaningful change in the educability of the culturally deprived will come through modifying and directing very early development of motivation, learning sets, attitudes and values.

The Ypsilanti Carnegie Infant Education Project attempts to modify the educability of culturally deprived children by working with the mother and child. At the last report (Lambie and Weikart, 1970) the project had been operating only one year, but interim results show the program to be effective. The authors state on p. 430:

> Perhaps the most important observation is that the process of a teacher, a mother, and an infant getting ready to learn together is even more critical than what is actually done. To be sure, the teacher must have ideas and "expertise" to assist the mother and infant in learning, but that is a long way from simply providing a family with a series of exercises.

There is little doubt that major determinants of learning style and ability are fixed in the individual's early life and that environment plays a dominant role. Mason (1970) provides a thoughtful discussion of the effects of environmental deprivation on learning. Many people concerned with education express the belief that, if successful, preschool education and training will aid in developing students with better dispositions and abilities for learning. Many student characteristics such as learning set and style, motivation, attitude, concept attainment, etc., which appear as given at school age, may be open to modification in preschool years.

SUMMARY

This section reviewed research on some of the important student characteristics that affect educational outcomes. A basic problem with education research is that it has generally failed to consider the interaction of individual student characteristics with instructional methods, teacher characteristics, and type of learning task.

Within psychology, the study of ability structure is complex and lengthy, and there is no general agreement about the level of specificity necessary to describe the structure nor is there agreement about the degree of genetic influence on abilities. Moreover, there is no conclusive evidence for an interaction between any special ability and education factors, although there are some indications that these interactions exist. There is, however, a clear indication that general intelligence is substantially correlated with ability to learn, especially for abstract and complex material; and in addition, there is strong evidence of an interaction between intelligence and education treatment (e.g., instruction, task). Such findings indicate that to be effective, educators must develop methods tailored for individual ability. Previous attempts to do this via ability-grouping failed because education programs were not developed that specifically fit the needs of the separate groups.

Other scattered research findings indicate that many factors differentiate students and their responses to specific education programs. For example, creativity is not highly dependent on intelligence (using the terms to define broad categories rather than unitary abilities), nor does high intelligence guarantee creativity; but there are indications that the creative person requires a different educational approach than the less creative individual. More generally, differences in level of concept attainment exist at school age and thus carry important implications for instruction design.

A number of personality variables (need for achievement, autonomy, and anxiety, among others) appear to influence school achievement and to interact with educational factors, but the evidence is not highly conclusive. The apparent importance of noncognitive factors on school achievement has led to a growing interest in the effect that preschool years have on educational outcomes. Findings from a number of experimental studies and preschool education programs support the assumption that major determinants of achievement are established in these years.

IMPLICATIONS

Research is making it increasingly clear that students respond differently to factors in education on the bases of their own individual characteristics. Research must continue to define student and educational characteristics and to find the most effective combinations, i.e., for a given set of student characteristics what are the best educational characteristics (given some set of objectives). Data systems must be designed to capture information on individual student characteristics so they can be related to performance data. To determine relevant student characteristics, research programs will require data on factors discussed in this section. Unfortunately, tests for most of these characteristics have not been adequately developed, although some promising ones are available.

The general areas in which data are needed both for research and for accountability include:

1. *General and Specific Abilities.* General Ability (Intelligence) is related to specific educational characteristics, and although the research has not as yet indicated any strong evidence for the effect of specific abilities, this area should be pursued. The lack of positive findings is likely due more to poor tests and faulty research than to an actual absence of the effect of these abilities.
2. *Level of Concept Attainment.* Concept Attainment is undoubtedly related to educational characteristics, although the experimental evidence from classroom studies is not highly positive. Laboratory studies indicate an individual characteristic in concept attainment, and educational programs should be aimed at such individuality. Tests of education readiness are included in this category even though the factors affecting readiness may be different from those affecting concept attainment. The complex issues of prior learning, maturation and genetics operate in both readiness and concept attainment.
3. *Creativity.* Creativity is certainly difficult to measure, but since it appears that the creative individual may have quite different reactions to educational characteristics from those of the less creative person, it is crucial to pursue the topic. Eventually an accountability system must be able to evaluate educational outcomes for the creative individual and assess the efficacy of a program in terms of his achievements.
4. *Personality.* Personality factors undoubtedly play a large role in determining the individual's unique reaction to characteristics of education programs, but attempts to measure personality factors have had little success; although some particular factors such as need-achievement and anxiety level show promise. Moderately consistent findings of a relationship between the high-achiever (as a personality characteristic) and instructional method have been reported. Data on personality factors should be part of a future data system, at least on a sample basis to allow further determination of how these factors influence educational outcomes.

VI. CONCLUSIONS AND DISCUSSION

In the beginning, the primary goal of this Report was to define, however loosely, the kinds of data that future school information systems would require for accountability. A secondary goal was to determine the kind of research needed to uncover factors relevant to refining accountability and to defining the data requirements for this research. A major effort was directed toward reviewing and summarizing education research findings. As this effort progressed, it became increasingly clear that the primary goal is largely unattainable at this time. Research indicates that data requirements for accountability are largely unknown because factors affecting educational outcomes are generally unknown. Program budgeting, cost-effectiveness analysis, or any other form of accountability is no better than the output measures that serve as criteria. If student achievement is the critical output measure, then it appears that accountability is in serious trouble because at present there is little the school can do to affect outcomes. Therefore, this Report has necessarily focused on what originally had been a secondary objective: to indicate what research must be done to better understand education and to make accountability possible.

Education research in the conventional classroom has failed to find clear and conclusive evidence regarding factors that affect student achievement. It is true that a number of socio-economic variables influence student achievement, but as yet there has been little success in modifying low achievement through innovation (Hellmuth, 1970). Not only has research on education generally failed, but little is known about the relationship between school expenditures and student achievement (Averch, Carroll, and Donaldson, 1971). There are two basic schools of thought giving reasons for lack of positive findings in education research. One view holds that the wrong data or faulty data are collected, and that if "good" data on relevant factors were analyzed, the effects of education would become apparent. This view holds, either explicitly or implicitly, that the school does affect learning, but that the effect cannot be measured. Another view holds that although there certainly are data problems, the essential difficulty is that factors actually affecting student achievement are excluded from (most) school programs, or they lie outside the school's jurisdiction. In this view, schools do not affect learning so there is nothing to measure.

Both views are probably partly correct. As the review of achievement tests indicates, data on student achievement needs improvement, but in addition the factors affecting this outcome have not been sufficiently identified. Years of education research seem to demonstrate that searching for broad generalizations about educational achievement is fruitless. There are some positive findings related to small aspects or subdivisions of the total education process, however, and with continued research it may be that a taxonomy will develop within which limited generalizations hold. This is demonstrated in interactions between students and various educational factors, such interactions appearing to exist between student type and instructional method; between student type and teacher type; and between teacher type and instructional method.

The clearest evidence is for student-method, and student-teacher interactions. The evidence for teacher-method interaction is weaker, perhaps because it is rarely studied. The student-method interaction implies that students require individualized programs with respect to instructional methods. The student-teacher interaction implies that students and teachers must be matched in terms of their ability to work together. Until schools develop programs allowing them to utilize these interactions, there may be little they can do to affect student achievement. Educational research has ignored individual differences, treating students and teachers as homogeneous bodies, and the results of past research may simply indicate that on the average (summed over students or teachers) school-controlled factors do not determine outcome.

Research on instructional method has yielded some promising results, especially the studies on organization and sequencing of instruction. In light of the evidence on student-method interaction, it appears that this approach can be easily adapted to individualized instruction. But most of the relevant instruction research has been done in the psychological laboratory, and the meaning of results for classroom learning is not clear. In any event, it seems that the implications of instructional method research are not as immediately important to education as are the findings on interaction. The expected difference in student achievement between relatively poor and excellent instructional techniques can probably not be nearly as great as the differences produced in utilizing the interaction phenomenon, particularly that between student and teacher. If a student in fact responds well to a particular teacher, then that teacher is obviously using, among other things, an instructional method that is compatible with that student.

Despite a tremendous number of educational studies, there is still no clear, comprehensive understanding of education. There are a number of explanations in addition to those mentioned above. Classroom studies often fall short because of poor measures and limited criteria of achievement, and because of statistical errors in the treatment of data. Studies often lack experimental control, and non-random factors confound or conceal elements in interpretation. Of course, it may be that the research model used in education research, based on physics and agricultural experimentation, does not fit. Perhaps a biological or legal (argumentation) model would be better.

At another level, research programs and achievement measures have not been

related to clearly defined education objectives, and noncognitive achievement has been almost totally ignored even though it is this kind of achievement that program objectives often seem to fit best. Occasionally one finds an outstanding study in terms of design, scope and analysis, and the results are often impressive, but there is rarely any follow-on to these studies so they stand unconfirmed. Finally, almost all classroom studies lack sufficient data on classroom transactions, so that it is impossible to determine what actually transpired in the learning situation.

Laboratory studies are generally better designed, and statistical treatment of data is more adequate. Many studies do contain glaring methodological errors, but their real shortcoming for understanding classroom learning lies in the fact that they employ learning tasks that are largely irrelevant to classroom learning. In addition, learning takes place under artificial conditions, making it difficult to extrapolate from the laboratory to the classroom.

In view of the current state of knowledge about the education process it seems highly possible that great emphasis on accountability may have a number of undesirable effects. If schools (teachers, etc.) are held accountable for student achievement (on reading and math tests or in other subject areas), then in all probability the school will find ways to produce high scores on these tests. At best, however, the tests measure only a small part of educational objectives and one cannot help but wonder what will happen to other kinds of achievement.

Moreover, the belief that the introduction of accountability will significantly alter student achievement must be based on the assumption that administrators and teachers know what to do in order to improve educational outcomes, but they are not sufficiently motivated to do so. There is little evidence to support this belief. It seems that administrators and teachers (and everyone else for that matter) know very little about factors affecting educational outcomes, and therefore they cannot change outcomes whether they want to or not. Before accountability can affect educational outcomes, teachers must have methods at their disposal that affect outcome.

It seems apparent that accountability, and education itself, need broad innovation. Additional research and real-world innovative approaches must determine the kinds of innovation necessary. Future research needs to determine the interactive factors among students, teachers, and instructional methods. By and large education research needs more continuity so that studies build on previous work; there is too much duplication of trivial research and too little replication of meaningful research. One of the most difficult problems that research must solve is the identification, measurement, and modification of noncognitive achievement. This needs to be developed in close conjunction with a look at the objectives of education and methods for evaluating these objectives.

Stating educational objectives is difficult, and it is even more difficult because objectives depend on personal values; not only do values change from person to person, but they are seldom overt. Stake (1970) addresses the problem of evaluating subjective data, and pleads that this much neglected area be considered in educational research and evaluation. Progress in educational research is highly dependent upon evaluating objectives and values. Moreover, it is important to consider

differences in values about educational objectives in different segments of the population because it is apparent that vast differences exist between various age groups and subcultures concerning the relative importance of various educational goals.

Appendix

INTERACTION EFFECTS

Some students do better with some instruction method, whereas other students perform better with other methods. This phenomenon is called an interaction, in this case, between student and instruction method. Suppose that two instructional methods (A_1 and A_2) are used in separate classrooms that have an equal number of students in each of two categories (B_1 and B_2—bright-dull, creative-noncreative, autonomous-dependent, etc.). At the end of a school year, the average achievement of each type of student under each instructional method is found. Figure 1 shows how an interaction would appear in the results. Since two types of students and two

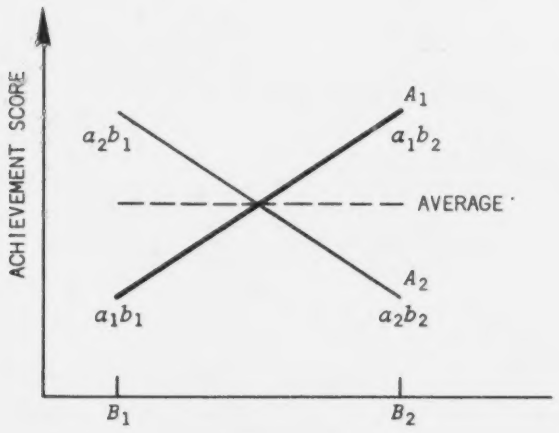

Fig. 1—Two-factor interaction between instructional methods (A_1 and A_2) and student type (B_1 and B_2) on achievement

instructional methods are used, the points (a_1 b_1, a_1 b_2, etc.) representing the mean score for that condition are connected by straight lines within methods. Students of type B_1 do better under method A_2 than under A_1, whereas the opposite is true for type B_2 students. Had the study been designed simply to determine the relative effectiveness of the instruction methods without also identifying the student types, no differences between methods would have been found, as shown in Fig. 1 by the dotted line, which indicates the average achievement for all students (type B_1 and B_2) within a method.

This example demonstrates only a two-factor interaction—student and method. Suppose that some teachers are better with some students and that both or either are better with some instructional method, leading to a three-factor interaction among student, teacher, and method, as shown in Fig. 2. In the figure, C_1 teachers

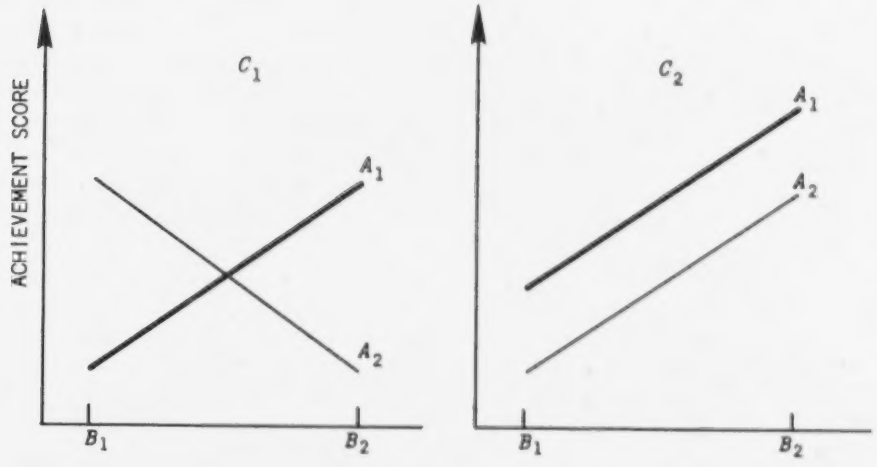

Fig. 2—Three-factor interaction between instructional method (A_1 and A_2), student type (B_1 and B_2) and teacher type (C_1 and C_2) on achievement

do well with B_1 students under method A_2, and with B_2 students under method A_1, and do poorly in the other cases. C_2 teachers are uniformly better with method A_2; in this case, B_1 students are uniformly better than B_2, which appears to be due to the teacher effect.

Interactions can be very complex, and can only be determined by appropriate experimental design. Many contradictory and inconclusive results in the past may have occurred because of unidentified interaction effects. For example, two teaching

methods may show significant differences because they happen (by chance) to have a predominantly homogeneous group of students and one method was better for such students. Another study of the same two methods may show no difference because students are not homogeneous with respect to some important characteristic and interaction effects conceal the true differences due to methods.

BIBLIOGRAPHY

Adelson, J., "Personality," *Annual Review of Psychology, 20,* 1969, pp. 217-252.

Allen, D. I., "Some Effects of Advance Organizers and Level of Questions on the Learning and Retention of Written Social Studies Material," *Journal of Educational Psychology, 61,* 1970, pp. 333-339.

Allen, W. H., "Instructional Media Research: Past, Present and Future," *Audio-Visual Communications Review, 29,* 1971, pp. 5-18.

Anastasi, A., "Psychology, Psychologists, and Psychological Testing," *American Psychologist,* 22, 1967, pp. 297-306.

Anderson, R. C., "Educational Psychology," *Annual Review of Psychology, 18,* 1967, pp. 103-164.

Angoff, W. H., "Scales, Norms, and Equivalent Scores," in R. L. Thorndike (ed.), *Educational Measurement,* American Council on Education, Washington, D.C., 1971, 508-600.

Ausubel, D. P., *The Psychology of Meaningful Verbal Learning,* Grune and Stratton, New York, 1963.

——, "An Evaluation of the Conceptual Scheme Approach to Science Curriculum Development," *Journal of Research in Scientific Teaching, 3,* 1965, pp. 255-264.

Averch, H., S. J. Carroll, and T. S. Donaldson, *What Do We Know About Educational Effectiveness?* Report to the President's Commission on School Finance, The Rand Corporation (to be published).

Bloom, B. S., T. J. Hastings, and G. F. Madaus, *Handbook on Formative and Summative Evaluation of Student Evaluation of Student Learning,* McGraw-Hill, New York, 1971.

Bormuth, J. R., *On the Theory of Achievement Test Items,* University of Chicago Press, Chicago, 1970.

Brophy, J. E., and T. L. Good, "Teacher's Communication of Differential Expectations for Children's Classroom Performance," *Journal of Educational Psychology, 61,* 1970, pp. 365-374.

Brownell, W. A., and A. G. Moser *Meaningful Versus Mechanical Learning: A Study in Grade Three Subtraction,* Duke University Research Studies in Education, No. 8, Durham, North Carolina, Duke University Press, 1949.

Burton, B. B., and R. A. Goldberg, *The Effect of Responses Characteristics in Multiple Life and Choice Alternatives on Learning During Programmed Instruction,* American Institute for Research, San Mateo, California, 1962.

Butler, A. L., *Current Research in Early Childhood Education,* E/K/N/E, NEA Center, Washington, D.C., 1970.

Chomsky, N., "Review of Verbal Behavior," *Language, 35,* 1959, pp. 26-58.

Chu, G. C., and W. Schramm, *Learning from Television,* Institute for Communications Research, Stanford University, 1967.

Coffman, W. E., Essay Examinations, *Educational Measurement,* R. L. Thorndike (ed.), American Council on Education, Washington, D.C., 1971, pp. 271-302.

Cohen, D. K., "Politics and Research: Evaluation of Social Action Programs in Education," *Review of Educational Research, 40,* 1970, pp. 213-238.

Corey, S. N., "The Nature of Instruction," in P. C. Lang (ed.), *Programmed Instruction,* University of Chicago Press, Chicago, 1967, pp. 5-27.

Cronbach, L. J., "The Two Disciplines of Scientific Psychology," *American Psychology, 12,* 1957, pp. 671-684.

——, "The Logic of Experiments on Discovery," in L. S. Shulman and E. R. Keisler (eds.), *Learning by Discovery: A Critical Appraisal,* Rand McNally and Company, Skokie, Illinois, 1966.

——, *Essentials of Psychological Testing,* Harper and Row, New York, 1970.

——, and L. Furby, "How Should We Measure 'Change'—or Should We?" *Psychology Bulletin, 74,* 1970, pp. 68-80.

Cronbach, L. J., and R. E. Snow, *Final Report: Individual Differences in Learning Ability as a Function of Instructional Variables,* Stanford University, Stanford, California, 1969.

Dahlstrom, W. G., "Personality," *Annual Review of Psychology, 21,* 1970, pp. 1-48.

Deese, J., "Behavior and Fact," *American Psychology, 24,* 1969, pp. 515-522.

Dellas, M., and E. L. Gaier, "Identification of Creativity: The Individual," *Psychology Bulletin, 73,* 1970, pp. 53-73.

Deneburg, V. H. (ed.), *Education of the Infant and Young Child,* Academic Press, New York, 1970.

Donaldson, T. S., *Subjective Scaling of Student Performance,* The Rand Corporation, P-4596, 1971.

Elkind, D., and A. Sameroff, "Developmental Psychology," *Annual Review of Psychology, 21,* 1970, pp. 191-238.

Engelmann, S., "The Effectiveness of Direct Instruction on IQ Performance and Achievement in Reading and Arithmetic," in J. Hellmuth (ed.), *Disadvantaged Child,* Vol. 3, *Compensatory Education: A National Debate,* Brunner/Mazel, New York, 1970.

Ferguson, G. A., "Human Abilities," *Annual Review of Psychology, 16,* 1965, pp. 39-61.

Flavel, J. H., and J. P. Hill, "Developmental Psychology," *Annual Review of Psychology, 20,* 1969, pp. 1-56.

Fleishman, E. A., and C. J. Bartlett, "Human Abilities," *Annual Review of Psychology, 20,* 1969, pp. 349-380.

Gagné, R. M., "The Acquisition of Knowledge," *Psychological Review, 69,* 1962, pp. 355-365.

—— (ed.), *Learning and Individual Differences,* Merrill, Columbus, Ohio, 1967.

——, "Contributions of Learning to Human Development," *Psychological Review, 75,* 1968, pp. 177-191.

——, and W. D. Rohwer, Jr., "Instructional Psychology," *Annual Review of Psychology, 20,* 1969, pp. 381-418.

Garrett, M., and J. A. Fodor, "Psychological Theories and Linguistic Constructs," in T. R. Dixon, and D. L. Horton, (eds.), *Verbal Behavior and General Behavior Theory,* Prentice-Hall, Englewood Cliffs, New Jersey, 1969, pp. 451-477.

Getzel, J. W., and P. W. Jackson, "The Teachers Personality and Characteristics," in N. L. Gage, (ed.), *Handbook of Research on Teaching,* Rand McNally, Chicago, 1963, pp. 506-582.

Gintis, H., "Education, Technology, and the Characteristics of Worker Productivity," *The American Economic Review, LXI,* 1971, pp. 266-279.

Glaser, R., and A. J. Nitko, "Measurement in Learning and Instruction," in R. L. Thorndike (ed.), *Educational Measurement,* Washington, D.C., American Council on Education, 1970.

Goodland, J. I., "Curriculum: State of the Field," *Review of Education Research, 39,* 1969, pp. 367-375.

Gotkin, L. G., and J. McSweeney, "Learning from Teaching Machines," in P. C. Lang (ed.), *Programming Instruction,* University of Chicago Press, Chicago, 1967, pp. 255-283.

Gough, H. G., and M. B. Fink, "Scholastic Achievement Among Students of Average Ability as Predicted from the California Psychological Inventory," *Psychology in the Schools, 1,* 1964, pp. 374-380.

Groteluesehin, A., and D. D. Sjorgren, "Effects of Differentially Structured Introductory Material and Learning Tasks on Learning and Transfer," *American Educational Research Journal, 2,* 1968, pp. 191-202.

Guilford, J. P., *The Nature of Human Intelligence,* McGraw-Hill, New York, 1967.

Hanley, E. M., "Review of Research Involving Applied Behavior Analysis in the Classroom," *Review of Educational Research, 40,* 1970, pp. 597-625.

Harris, A. J., "The Effective Reading Teaching," *The Reading Teacher, 23,* 1969, pp. 195-204.

Harris, C. W. (ed.), *Problems in Measuring Change,* University of Wisconsin Press, Madison, Wisconsin, 1963.

Hartup, W. W., and A. Yonas, "Developmental Psychology," *Annual Review of Psychology, 22,* 1971, pp. 169-392.

Heckhausen, H., *The Anatomy of Achievement Motivation,* Academic Press, New York, 1967.

Hellmuth, J. (ed.), *Disadvantaged Child,* Vol. 3, *Compensatory Education: A National Debate,* Brunner/Mazel, New York, 1970.

Heron, M. D., "The Nature of Scientific Enquiry as Seen by Selected Philosophers, Science Teachers, and Recent Curricular Materials," Ph.D. Dessertation, Chicago, University of Chicago, 1969.

Hoepfner, Ralph (ed.), *CSE Elementary School Test Evaluation,* Center for the Study of Evaluation, UCLA Graduate School of Education, Los Angeles, 1970.

Holtzman, W. H., "The Changing World of Mental Measurement and Its Social Significance," *American Psychologist, 26,* 1971, pp. 546-553.

Jensen, A. R., "How Much Can We Boost IQ and Scholastic Achievement?" *Harvard Educational Review, 39,* 1969, pp. 1-123.

——, "Another Look at Culture Fair Testing," in J. Hellmuth (ed.), *Disadvantaged Child,* Vol. 3, *Compensatory Education: A National Debate,* Brunner/Mazel, New York, 1970, pp. 53-101.

Kagan, J., "On Class Differences and Early Development," in V. H. Denenberg (ed.), *Education of the Infant and Young Child,* Academic Press, New York, 1970, pp. 5-24.

Karraker, R. J., and L. A. Doke, "Errorless Discrimination of Alphabet Letters: Effects of Time and Method of Introducing Competing Stimuli," *Journal of Experimental Education, 38,* 1970, pp. 27-35.

Klein, G. S., H. L. Barr, and D. L. Wolitzky, "Personality," *Annual Review of Psychology, 18,* 1967, pp. 465-560.

Klein, S. P., "The Uses and Limitations of Standardized Tests in Meeting the Demands for Accountability," Center for the Study of Evaluation, UCLA Evaluation Comment, 2, No. 4, January 1971.

Kohlberg, L., "Early Education: A Cognitive-Developmental View," *Child Development, 39,* 1968, pp. 1013-1962.

Lambie, D. J., and D. P. Weikart, "Ypsilanti Carnegie Infant Education Project," in J. Hellmuth (ed.), *Disadvantaged Child,* Vol. 3, *Compensatory Education: A National Debate,* Brunner/Mazel, New York, 1970.

Lang, P. C. (ed.), *Programmed Instruction,* 66th Yearbook of the National Society for the Study of Education, Part II, University of Chicago Press, Chicago, 1967.

Lernon, R. T., *Accountability and Performance Contracting,* invited address to the American Educational Research Association, New York City, February 5, 1971.

Maier, M., and P. B. Jacobs, "The Effects of Variations in a Self-Instructional Program on Instructional Outcomes," *Psychology Reports, 18,* 1966, pp. 539-546.

Mason, W. A., "Early Deprivation in Biological Perspective," in V. H. Dennenberg (ed.), *Education of the Infant and Young Child,* Academic Press, New York, 1970.

Merrill, M. D., "Correction and Review on Successive Parts in Becoming a Hierarchical Task," *Journal of Educational Psychology, 56,* 1965, pp. 225-234.

——, K. Barton, and L. E. Wood, "Specific Review in Learning a Hierarchical Imaginary Science," *Journal of Educational Psychology, 61,* 1970, pp. 102-109.

Merrill, M. D., and L. M. Stolurow, "Hierarchical Preview Versus Problem Oriented Review in Becoming an Imaginary Science," *Journal of American Educational Research, 3,* 1966, pp. 251-261.

Nitko, A. J., *A Model for Criterion-Referenced Tests Based on Use,* Paper presented at the Annual Meeting of the American Educational Research Association, New York, February 4-7, 1971.

Piland, J. C., and E. A. Lemke, he Effect of Ability Grouping on Concept Learning," *Journal of Educational Research, 64,* 1971, pp. 209-212.

Pressey, S. J., "Teaching Machines (and Learning Theory) Crises," *Journal of Applied Psychology, 47,* 1963, pp. 1-6.

Rist, R. C., "Student Social Class and Teacher Expectations: The Self-Fulfilling Prophecy in Ghetto Education," *Harvard Educational Review, 40,* 1970, pp. 411-451.

Rohwer, W. D., Jr., M. S. Ammon, N. Suzuki, and J. R. Levin, "Population Differences and Learning Proficiency," *Journal of Educational Psychology, 62,* 1971, pp. 1-14.

Romberg, T. A., "Current Research in Mathematics Education," *Review of Education Research, 39,* 1969, pp. 473-492.

Rosenshine, B., "Evaluation of Instruction," *Review of Educational Research, 40,* 1970a, pp. 279-300.

——, "The Stability of Teacher Effects Upon Student Achievement," *Review of Educational Research, 40,* 1970b, pp. 647-662.

——, and J. Furst "Current and Future Research on Teacher Performance Criteria," in B. W. Smith (ed.), *Research on Teacher Education: A Symposium,* Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

Rosenthal, R., and L. Jacobson, *Pygmalion in the Classroom: Teacher Expectation and Pupils' Intellectual Development,* Holt, Rinehart, & Winston, New York, 1968.

Ryder, R. G., "Birth to Maturity Revisited: A Canonical Reanalysis," *Journal of Personality and Social Psychology, 7,* 1967, pp. 168-172.

Saettler, P., *A History of Instructional Technology,* McGraw-Hill Book Company, New York, 1968.

Samuels, S. J., "Effects of Pictures on Learning to Read, Comprehension and Attitudes," *Review of Educational Research, 40,* 1970, pp. 397-407.

Sarason, I. G., and R. E. Smith, "Personality," *Annual Review of Psychology, 22,* 1971, pp. 393-446.

Schwartz, P. A., Prediction Instruments for Educational Outcomes, *Educational Measurement,* R. L. Thorndike (ed.), American Council on Education, Washington, D.C., 1971, pp. 303-331.

Shulman, L. S., and E. R. Keislar (eds.), *Learning by Discovery: A Critical Appraisal,* Rand McNally and Company, Skokie, Illinois, 1966.

Shulman, L. S., "Reconstruction of Educational Research," *Review of Educational Research, 40,* 1970, pp. 371-396.

Skinner, B. G., *The Technology of Teaching,* Appleton-Century Crofts, New York, 1968.

Smith, H. A., Curriculum Development and Instructional Materials, *Review of Educational Research, 39,* 1969, pp. 397-414.

Snow, R. E., "Mental Abilities," *The Encyclopedia of Education* (in press, 1971).

——, "Unfinished Pygmalion," *Contemporary Psych· · 14,* 1969, pp. 197-200.

Stake, R. E., "Objectives, Priorities, and Other Judgm· . Data," *Review of Educational Research, 40,* 1970, pp. 181-212.

——, "Testing Hazards in Performance Contracting," Phi Delta Kappan, 12, 1971, pp. 583-589.

Stephens, J. M., *The Process of Schooling,* Holt, Rinehart and Winston, New York, 1967.

Stodolsky, S. S., and G. Lesser, "Learning Patterns in the Disadvantaged," *Harvard Review of Education, 37,* 1967, pp. 546-553.

Thelen, H. A., "Programmed Materials Today: Critique and Proposal," *The Elementary School Review, 64,* 1963a, pp. 189-196.

——, "Programmed Instruction: Insight vs. Conditioning," *Education, 83,* 1963b, pp. 416-420.

——, *Classroom Grouping for Teachability,* John Wiley & Sons, New York, 1967.

Turner, R. L., and D. A. Denny, "Teacher Characteristics, Teacher Behavior, and Changes in Pupil Creativity," *Elementary School Journal, 62,* 1969, pp. 265-270.

Vandenberg, S. G., "Contributions of Twin Research to Psychology," *Psychology Bulletin, 66,* 1966, pp. 327-352.

Vernon, P. E., "Ability Factors and Environmental Influence," *American Psychology, 20,* 1965, pp. 723-733.

Welch, W. W., "Curriculum Evaluation," *Review of Educational Research, 39,* 1969, pp. 429-444.

Westbury, I., "Curriculum Evaluation," *Review of Educational Research, 40,* 1970, pp. 239-260.

Wiggens, J. S., "Personality Structure," *Annual Review of Psychology 19,* 1968, pp. 293-350.